arXiv:2611.29444v1 [cs.LG] 22 Nov 2026

RumiFold：教大语言模型原生折叠蛋白质并推理其功能

Kevin Lu
东北大学
lu.kev@northeastern.edu

David Bau*东北大学
d.bau@northeastern.edu

Wengong Jin*东北大学
w.jin@northeastern.edu

摘要

我们能否教会通用语言模型折叠蛋白质？如果可以，那么学会折叠除了结构预测本身之外，还能带来什么？我们介绍了 RumiFold，这是一种预训练聊天模型，能够直接根据自身的隐藏状态预测蛋白质结构。RumiFold 通过多轮折叠处理蛋白质，随后利用这些折叠后的表示以自然语言回答问题。它达到了与 SimpleFold 等专用单表示模型相当的折叠准确率，同时在基因本体蛋白质功能预测和自由文本功能描述方面取得了最先进的性能。我们发现，折叠预训练会围绕蛋白质功能组织模型的隐藏状态，形成一个具有生物学基础的内部逐残基工作空间。因果干预表明，当 RumiFold 预测功能时，这一折叠工作空间能够连接结构与语言，为生物分子推理提供丰富的表征。我们将 RumiFold 视为迈向原生生物分子智能的一步：在这一愿景中，基于蛋白质的计算与基于语言的推理将成为通用语言模型相互融合的能力。代码、数据和模型见：rumifold.baulab.info。

1 引言

大语言模型能学会折叠蛋白质吗？如果能，除了预测结构之外，它还能获得什么？大语言模型已经可以调用 AlphaFold3 [Abramson et al., 2024] 来获取蛋白质结构，但这样做并不能教会它如何折叠。通过亲自学会执行这项生物分子计算，语言模型可以构建一个内部工作空间，从而更深入地推理蛋白质结构和功能。

当前将通用语言模型与蛋白质相结合的工作，大多将生物分子计算置于模型自身的表示之外，因此尚未充分利用这一机会。一种路径依赖围绕外部预测器和生物软件的工具调用与智能体工作流 [Huang et al., 2025年, Ghafarollahi and Buehler, 2024年, Swanson et al., 2025], , 在这种方式下，生物计算发生在工具内部，迫使语言模型根据有损的文本摘要进行推理而非更具表达力的内部表征。另一种路径是将序列或结构编码器桥接到文本大语言模型 [Fei 等, 2025，Fallahpour 等人，2026，Liu 等，2024，Wang 等，2025b]，依靠静态特征或预计算预测结果来生成描述，而不是构建蛋白质的内部表征以进行推理。当前缺少的是一种通用聊天模型，它能在自身的隐藏状态中折叠蛋白质，并基于结果进行推理，而不是读取其他模型生成的结构或预测结果。

我们介绍了 RumiFold，这是一种通用语言模型，经过训练可根据自身的隐藏状态预测蛋白质结构。我们使用 LoRA [Hu et al., 2021], 对 Qwen3-8B 进行微调，并配合冻结的 ESM-C 编码器 [ESM Team, 2024] 为每个残基提供输入嵌入。之后的所有处理都由

*共同顾问和共同通讯作者。

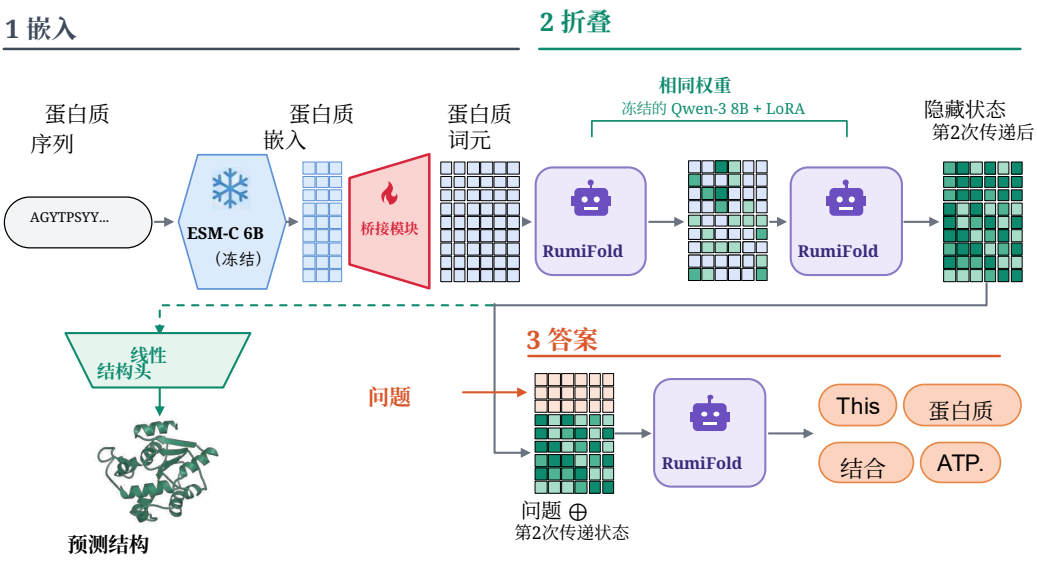


图 1：**RumiFold 对蛋白质进行嵌入，将其折叠，然后根据所学内容作答。***(1)* 嵌入：冻结的 ESM-C 编码器为每个残基生成嵌入，训练后的桥接模块将这些嵌入映射到语言模型的隐空间。*(2)* 折叠：语言模型（冻结的 Qwen3-8B，搭配 LoRA适配器）通过两轮折叠过程处理蛋白质，并将第一轮的隐藏状态传入第二轮，以改进其对蛋白质结构的表征。线性结构头根据生成的隐藏状态预测主链坐标。*(3)* 答案：该语言模型在以自然语言回答问题时，会关注这些折叠表示。

语言模型自身的层完成：它通过多轮折叠过程处理蛋白质，然后在用自然语言回答问题时关注由此产生的表示（图 1）。

贡献：我们发现了什么？

- **通用语言模型能够学会折叠蛋白质**。在冻结权重、仅训练低秩适配器和线性输出头的情况下，Qwen3-8B 折叠蛋白质的表现与小型专用单表征模型不相上下（FoldBench 单体测试：RumiFold 的 TM-score 为 0.840，SimpleFold 700M 为 0.832）（第 4 节）。
- **学会折叠有助于获得理解生物学功能所需的表征**。折叠预训练使模型的隐藏状态围绕蛋白质结构和功能重新组织。因此，RumiFold 在基因本体蛋白质功能预测和自由文本功能描述方面取得了最先进的表现，并且优于未经过折叠训练的匹配模型，以及通过读取 ESMFold2 隐藏状态进行预测的模型（RumiFold 的 wF1 为 0.683，GO-GPT v2 为 0.666；RumiFold 的 BLEU-4 为 41.0，Prot2Text-V2 为 39.0）（第 5 节）。
- **该模型通过折叠构建内部工作空间中的表征**。RumiFold 的折叠轮次构建逐残基工作空间，语言模型随后利用该工作空间回答问题。通过因果干预，我们表明，其功能预测依赖于这一工作空间，并且底层表示捕捉到了具有功能意义的特征和残基（例如近期在人体 TMEM68 中鉴定出的功能性残基 H130）（第 6 节）。

迈向原生生物分子智能。语言模型之所以具备广泛能力，是因为它们亲自处理了种类繁多的问题：预测、解释、翻译和求解，最终将推理和规划内化为模型的一部分。要具备生物学方面的能力，或许也需要采用同样的方法：让模型练习科学技能，直到它学会相关计算本身，并能够运用所学进行推理。我们认为，RumiFold 是迈向原生生物分子智能的一步，使蛋白质计算和基于语言的推理成为通用语言模型整合的能力。

2 相关工作

将蛋白质编码器与大语言模型配对。大多数近期的功能预测器会将蛋白质编码器的输出转换为大语言模型可读取的词元，再让大语言模型撰写答案。GO-GPT 根据 ESM-2 特征解码出基因本体术语；BioReason-Pro [Fallahpour 等人 2026], 提出的这一方法，则为推理大语言模型添加了基因本体、InterPro 和相互作用上下文；Prot2Text-V2 [Fei 等人 2025], 、Evolla [Zhou 等人 2025], 、BioMedGPT [Luo 等人 2023], 、EvoLlama [Liu 等人 2024] 以及 Prot2Chat [Wang 等人，2025b]则将序列或结构编码器连接到 Llama 系列模型。另一类方法会将序列以纯文本形式呈现 [Fang 等，2023，Wang 等，2023b，Shen 等人， 2024]。RumiFold 有一个关键区别：首先由语言模型本身处理蛋白质，并通过训练使其完成折叠，然后才向模型提问。

这些模型是否真正利用了蛋白质？ Wu 等人 [2025] 表明，在基于文本的基准测试中，检索相似的训练蛋白质，其表现可与经过训练的蛋白质大语言模型媲美，甚至更胜一筹；Fang 等人 [2026] 则通过输入扰动和证据冲突测试表明，一些模型依赖的是随附文本，而非生物学输入。我们采纳这两点经验：所有功能结果均按其与训练集的序列一致性或新颖性进行分层，并且我们会检验模型是否以可解释的方式利用蛋白质，而不只依赖准确率数值。

结构预测。 AlphaFold2 [Jumper 等人2021], 的出现带来了高准确度的蛋白质结构预测。该方法读取多序列比对，并在多轮迭代中不断细化蛋白质的成对表示。随后出现了两种简化方案。蛋白质语言模型被用作 BERT 风格编码器，同时也能捕捉 MSA 提供的进化信息：ESMFold [Lin 等人2023] 将折叠主干网络（仍采用成对表示）连接到 ESM-2，并根据单条序列预测结构。较新的一类方法则进一步舍弃成对表示，仅根据逐残基序列表示预测结构，其中包括 ESM3 [Hayes 等人 2025], 、DPLM-2 [Wang 等人 2024] 以及 SimpleFold [Wang 等人，2025a]。RumiFold 属于后一类，但有一点不同：它的主干网络是通用语言模型，而非针对蛋白质从头训练的网络。

3 RumiFold

3.1 架构（图1）

输入。 每条输入蛋白质链都由冻结的 ESM-C-6B 编码器进行嵌入 [ESM 团队， 2024]；我们取每个残基 i在最后一层的状态$\mathbf{e}_i \in \mathbb{R}^{2560}$。桥接模块将其映射到 Qwen 的嵌入空间，得到向量$\mathbf{h}_i^0 \in \mathbb{R}^d$，并在位置 i处替代词元嵌入输入主干网络：

$$\mathbf{h}_i^0 = \lambda \,\mathrm{LayerNorm}\big(W\,\mathrm{LayerNorm}(\mathbf{e}_i) + \mathbf{b}\big), \qquad \lambda = \sqrt{\frac{1}{|V|\,d}\sum_{v=1}^{|V|}\big\|\mathbf{E}_v\big\|_2^2}, \tag{1}$$

其中， $W \in \mathbb{R}^{d\times 2560}$ 和$\mathbf{b}$ 是可学习的， $d = 4{,}096$ 是 Qwen 的隐藏层宽度，$\mathbf{E}_v$ 是 Qwen 对词汇标记 v的嵌入。常数 λ 是 Qwen 词元嵌入矩阵各项的均方根，因此残基以普通词元的尺度进入主干网络。RumiFold 接受一条或两条链。其他架构细节见附录 A。

主干网络。 主干网络采用 Qwen3-8B [Yang 等人， 2025]。全部 36 个块均被冻结，并在每个块的查询、键、值和输出投影以及三个 MLP 投影（共 252 个投影）上放置秩为 256 的 LoRA 适配器，总参数量约为 7 亿。

三次传递，两种注意力掩码。RumiFold 在蛋白质上运行其主干网络三次，使用两种不同的注意力掩码（第 A.1 节中的图9）。同一组 LoRA 权重用于全部三次传递。

折叠轮次。 前两轮只输入蛋白质，并移除因果掩码：每个残基都关注每条链上的每个残基。令$\mathbf{s}_i^{(1)}$ 表示第一轮结束后残基 i 的最终隐藏状态。第二轮接收

$$\mathbf{h}_i^{\mathrm{in}} = \mathbf{h}_i^0 + \lambda\,\tanh(g)\,\mathrm{LayerNorm}\big(\mathbf{s}_i^{(1)}\big), \tag{2}$$

与 AlphaFold2 [Jumper 等人 2021], 的循环机制相同，其中$\mathbf{h}_i^0$和 λ 是方程（1）中的量，而g 是一个初始化为零的可学习标量，因此训练从一个不进行循环的模型开始。结构由第二轮读取。

问题轮次。 第三轮以因果方式处理文本，正如语言模型通常所做的那样；蛋白质作为前缀，模型可以关注它，但不能对其进行更改。设 $\mathbf{k}_j^\ell$ 和 $\mathbf{v}_j^\ell$ 为第 ℓ 层在第二个折叠轮次中为残基 j 计算得到的键和值。在第 ℓ 层，对于带有查询 $\mathbf{q}_t^\ell$ 的文本词元 t ，注意力输出为

$$\mathbf{o}_t^\ell = \sum_{j=1}^{L} \alpha_{tj}^\ell \, \mathbf{v}_j^\ell + \sum_{u \le t} \alpha_{tu}^\ell \, \tilde{\mathbf{v}}_u^\ell, \qquad \alpha_{t\cdot}^\ell = \mathrm{softmax}\big(\mathbf{q}_t^\ell \cdot [\,\mathbf{k}_{1:L}^\ell,\, \tilde{\mathbf{k}}_{1:t}^\ell\,]\big), \tag{3}$$

其中，波浪号表示文本自身的键和值。这是一个普通的键值缓存，但与语言模型为其自身提示构建的缓存有两点不同：它是在双向注意力下计算得到的，因此每个残基的条目已经反映了整个蛋白质；并且它保存了折叠状态的全部 36 层，而不是每个残基仅有一个输入嵌入。文本位置紧接在蛋白质的位置之后。信息不会反向流动：折叠轮次从不关注问题，因此预测结构不会取决于所提问题。

从隐藏状态到结构。 蛋白质主链是一条由残基组成的链，每个残基贡献三个原子（N、Cα、C），这些原子具有几乎固定的局部排列。因此，预测主链相当于确定每个残基的位置 以及朝向：一个刚性坐标系 $T_i = (R_i, \mathbf{t}_i)$，其中包含一个旋转和一个位置。在第二个折叠轮次的最终隐藏状态 $\mathbf{s}_i^{(2)}$ 经过层归一化后，再通过一个线性层处理；该线性层为每个残基输出九个数：位置$\mathbf{t}_i$ 以及两个经过正交归一化处理、从而得到 R_i 的向量。随后，将三个主链原子通过把固定的理想化模板 $\mathbf{u}_a$ 移入该坐标系来放置，即$\mathbf{p}_{i,a} = R_i \, \mathbf{u}_a + \mathbf{t}_i$ 分别对应 $a \in$ {N、 Cα、C}。整个注意力头约有 4.5 万个参数，而这个 80 亿参数语言模型经过微调的参数约有 7 亿个。这里没有成对表示、没有距离图，也没有扩散：结构是对语言模型隐藏状态进行线性读出的结果。

训练目标。 我们采用标准的结构损失。主要术语是 AlphaFold 的帧对齐点误差（FAPE）2 [Jumper 等人， 2021], 它在每个残基的局部坐标系中比较预测的原子位置与真实原子位置，因此不受结构旋转或平移的影响。我们分别计算链内和链间的误差，并加入两个较小的损失项，分别惩罚相邻残基之间不合理的键长以及原子间的碰撞。具体形式和权重见 A.2节。

3.2 训练

折叠数据。 我们使用 PDB 生物学组装体和 AlphaFold 数据库中的模型进行训练。对于 PDB 数据集，我们将发布日期截止时间设为 2022年1月1日，并选取通过 X射线晶体学或冷冻电子显微镜测定、分辨率为 3.5埃或更高且每条链至少含有 30 个残基的结构。对于 AlphaFold 数据库数据集，我们选取全局 pLDDT 至少为 70 的模型。两个来源合计提供了约 160 万个组装体，每个组装体的所有链合计最多包含 2,048 个残基。

两个模型。 我们报告了两个训练模型。*RumiFold* 在整个训练过程中持续进行语言回放，并按照以下阶段进行训练；它负责回答问题，我们也会评估其折叠能力。*RumiFold-Struct* 仅用于折叠训练，共进行了 176k 次更新；在折叠基准测试（第 4 节）和注意力头分析（D 节）中，我们将其用作仅结构参考。

阶段。 RumiFold 按照图2所示的四个阶段进行训练。这些阶段是累积式的：每个阶段都会增加一项任务，同时继续执行前面的任务，因此从第一次更新到最后一次更新，模型始终在进行折叠训练。让折叠对语言有用的关键步骤是第 2 阶段。在这一阶段，结构问题的答案（两个残基之间的距离、某个残基被埋藏的程度、哪些残基彼此接触）是在模型自身预测的坐标上计算的，而非依据天然结构。因此，模型不会因为复述有关蛋白质的事实而获得奖励；它得到奖励，是因为报告了其折叠结果所显示的信息。各阶段的切换取决于性能，而非固定的时间安排：当留出折叠准确率达到约 0.7 lDDT 时，便开始结构问答；当准确率

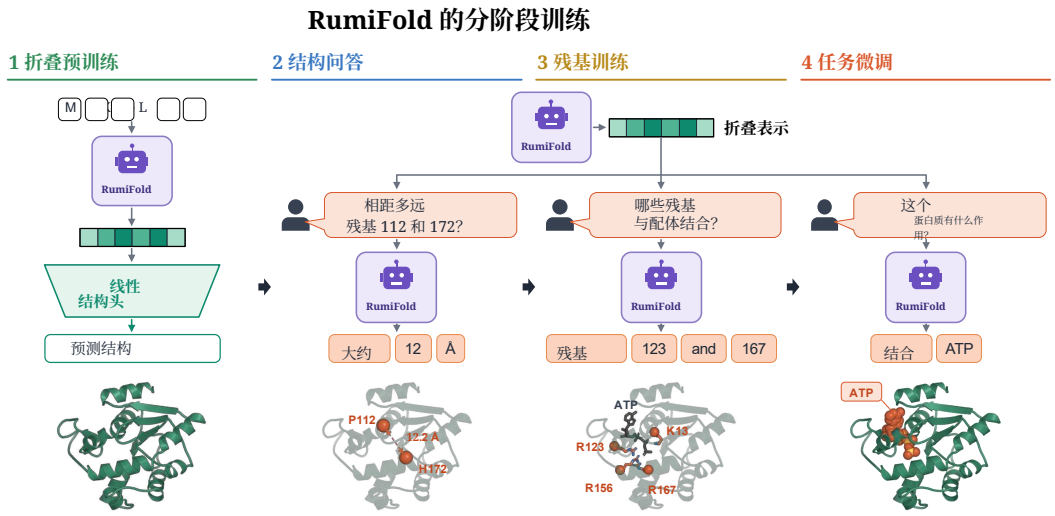


图2：**RumiFold 通过四个训练阶段学习折叠、理解和描述蛋白质**。第 *1* 阶段（折叠，更新次数 *0–9k*）：模型学习预测蛋白质结构，并在其隐藏状态中形成结构表征。第 *2* 阶段（结构问答，*9k–41k*）：模型学习调用这些表征，以自然语言回答有关其自身预测结构的问题。第 *3* 阶段（功能训练，*41k–95k*）：模型学习将结构表征与生物学功能联系起来。第 *4* 阶段（任务微调）：模型针对下游任务进行适配，包括功能预测、残基注释和蛋白质问答。这些阶段是累积的：每个阶段都会在此前阶段的任务基础上增加一项任务，因此同一个语言模型在全部四个阶段中都会持续进行折叠训练。

达到约 0.8 且模型能够可靠地回答结构问题后，便开始功能训练；大约完成一轮功能训练数据训练后，再进行任务微调。第 2 阶段后期使用的问题类型和受扰动的折叠结构详见附录 B，第 3 阶段的答案格式详见附录 H，训练数据及其留出集详见附录 A.3。

保留语言能力。回放提示取自 Tulu-3 混合数据集 [Lambert 等人， 2024],并由基础版 Qwen3-8B 重新生成回答（823k 条对话）。在这些轮次中，模型通过一个朝向基础版 Qwen 下一词元分布的 KL 项以及一个较小的交叉熵项进行训练；其中一半轮次的上下文中包含蛋白质。

计算。RumiFold 的第 1 至第 3 阶段在 8 块 A100 GPU 上训练了约三周。第 4 阶段针对每项任务进行微调（基因本体、Prot2Text），在 4 块 H100 GPU 上各耗时约十天。优化详情见 A.2节。

4 大语言模型学会折叠了吗?

我们评估 RumiFold 本身（即用于回答问题的模型），并将其与三个单表征模型 ESM3-open [Hayes 等人， 2025], DPLM-2 3B [Wang 等人， 2024]以及 SimpleFold 700M [Wang 等人，2025a], 进行比较，同时也与 ESMFold [Lin 等人， 2023]进行比较。我们使用 FoldBench 的 334 个单体目标 [Xu 等人， 2025] 以及 CASP16 的 155 个评估结构域[Kryshtafovych 等人， 2025], ，并在 Cα 坐标上计算 TM-score、lDDT 和 GDT-TS。

在 FoldBench 上，RumiFold 的 TM-score 达到 0.840，高于 SimpleFold 700M（0.832）、ESM3-open（0.783）和 DPLM-2 3B（0.709；图3）。在 CASP16 上，RumiFold 的得分为 0.714，而 SimpleFold 700M 为 0.706。RumiFold-Struct 将全部容量都用于结构预测，表现又略胜一筹（分别为 0.846 和 0.720）。SimpleFold 700M 的局部准确度更高：在两组数据上的 lDDT 均高出 0.02（分别为 0.802 对 0.781、0.730 对 0.706），GDT-TS 则高出 0.01。因此，RumiFold 对链的定位能力不逊于训练规模相近的专用单表征模型，但在局部细节上略逊一筹。参数量为 3B 的最大版 SimpleFold，在 FoldBench 上的 TM-score 与 RumiFold 相同，在 CASP16 上则落后 0.01（表 6）。我们还在

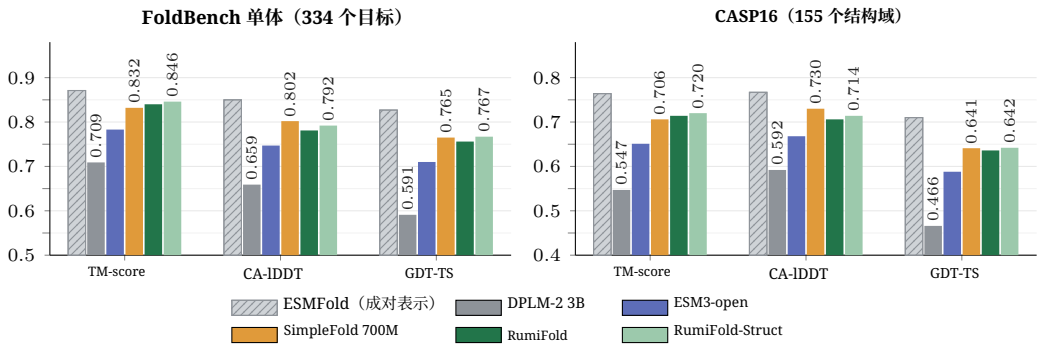


图3：**带适配器的冻结语言模型，其折叠效果与专用的单表征模型大致相当**。在 FoldBench 单体和 CASP16 结构域上的单序列结构预测：Cα TM-score、Cα-lDDT 和 GDT-TS。RumiFold 是同时能够回答问题的模型；RumiFold-Struct 仅针对折叠进行训练。DPLM-2、ESM3 和 SimpleFold 是单表征模型。ESMFold（阴影标记）使用成对表示，作为参考模型列出。

CAMEO22 和 CASP14 上进行了评估；这两个数据集在训练时按 30% 的序列一致性阈值筛除。我们在 C节报告了这些留出测试结果；在这两项测试中，SimpleFold 700M 均领先于 RumiFold。由于 RumiFold 可接收两条链作为输入，因此也能预测复合物，而其他单表征模型无法做到这一点（C节）。我们还通过比较预训练、打乱和随机初始化的权重（C.1节），探究语言模型的文本预训练是否有助于折叠，还是说任何同等规模的 Transformer 都能胜任。我们发现，在基础模型冻结、仅训练适配器的设置下，语言预训练确实有帮助。在训练后的模型中，折叠功能由一小部分追踪残基接触的注意力头承担；这些注意力头与基础版 Qwen 中用于匹配重复词元的注意力头有所重叠（D节）。

5 折叠知识是否有助于预测功能?

现在，我们转向构建折叠阶段所要解决的问题。能够进行折叠的模型不一定学会了可用于其他任务的知识，因此我们在两个功能基准测试上评估 RumiFold：基因本体预测和自由文本描述。在每个基准测试中，大量测试蛋白质在训练集中都有近缘蛋白质；对于这些蛋白质，召回率已足够。因此，我们根据测试蛋白质与训练集之间的序列一致性或新颖性报告结果。

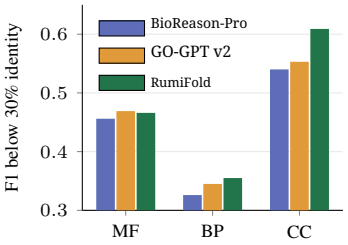
5.1 基因本体预测

设置。 我们从序列预测基因本体术语 [Ashburner 等人提出的 2000]分子功能、生物过程和细胞组分，所使用的基准测试是随 GO-GPT v2 发布的[Fallahpour 等人提出的 2026]。我们的所有模型都在其训练集（169,810 个带标签蛋白质）上进行训练，并在其蛋白质测试集上，使用其一级标签和信息加权 F1 [Zhou 等人提出的 2019]进行评估。我们报告其中两个模型。*RumiFold* 在该基准测试上完成功能训练和微调后，将其 GO术语以文本形式输出。*RumiFold-Struct + GO 头* 是仅使用结构的模型，经微调后配备了一个分类头，该分类头只读取语言模型的隐藏状态。两者在功能训练期间都保留较小的折叠损失，并且之后仍然能够折叠：在 FoldBench 上，功能微调使 RumiFold 的 TM-score 保持不变（$+0.001$、$p = 0.65$），而使后者的 TM-score 降低了 0.007。

结果。 RumiFold 的加权平均 F1 值达到 0.683，带有 GO 头的 RumiFold-Struct 达到 0.657；相比之下，已发布的 GO-GPT v2 检查点为 0.666，而 BioReason-Pro 使用该检查点的预测结果时为 0.655（表 1）；与 GO-GPT v2 的差异具有显著性（$p = 0.0003$，对测试蛋白质进行配对置换检验）。与已发布的 GO-GPT v2 相比，RumiFold 在分子功能（0.685 对 0.665，$p = 0.020$）和细胞组分（0.748 对 0.718，$p = 10^{-4}$）方面表现更好，在生物过程方面也略胜一筹（0.615 对 0.614）。RumiFold 只需在该基准测试上进行少得多的训练便取得了这一成绩：已发布的 GO-GPT v2 检查点已在训练集上训练 90 轮，而

表 1：**GO-GPT v2 基准测试上的基因本体预测**。在蛋白质测试集（3,726 个蛋白质）上，评估分子功能（MF）、生物过程（BP）和细胞组分（CC）的信息加权 F1，所有数据行均使用同一脚本评分。所有生成的答案，无论是我们的方法还是基线，均采用贪心解码：每个蛋白质生成一个答案，不进行采样，也不从候选项中选择。BioReason-Pro 会在提示中读取已发布的 GO-GPT v2 的预测结果。粗体表示最佳模型，下划线表示次佳模型。 *<50%*：对与任何训练蛋白质的序列一致性均低于 50% 的测试蛋白质取均值。右图：对与任何训练蛋白质的序列一致性均低于 30% 的 181 个测试蛋白质，按方面显示相同评分（BioReason-Pro RL、GO-GPT v2 和 RumiFold；坐标轴从 0.3 开始）。

模型	MF	BP	CC	Mean	<50%
Qwen3-8B，未训练	0.111	0.058	0.161	0.110	0.110
BioReason-Pro（监督微调）	0.617	0.571	0.679	0.622	<u>0.511</u>
BioReason-Pro（强化学习）	0.649	0.600	0.714	0.655	0.496
GO-GPT v2（已发布）	0.665	<u>0.614</u>	<u>0.718</u>	<u>0.666</u>	<u>0.511</u>
Qwen + ESM-C（无折叠预训练）	0.664	0.537	0.712	0.638	0.480
Qwen + ESMFold2 状态	0.633	0.577	0.705	0.638	0.431
RumiFold-Struct + GO 头	<u>0.675</u>	0.590	0.706	0.657	**0.521**
RumiFold	**0.685**	**0.615**	**0.748**	**0.683**	0.508



RumiFold 在完成折叠和功能训练阶段后训练了 16 轮。将未训练的 Qwen3-8B 的输入设为文本形式的序列时，其得分为 0.110，因此这些成果并非来自语言模型的先验知识。

已发表系统的增益。 两点观察有助于理解这些基线。首先，GO-GPT v2 在生物过程上的准确率在训练后期上升（第 25、50 和 75 个训练轮次分别为 0.432、0.504 和 0.592），其按序列同一性划分的表现显示，这些增益出现在最高同一性区间，这正是召回率不断累积的特征。其次，BioReason-Pro 的准确率与其提示中提供的 GO-GPT 预测结果一致，并未超过这些预测结果：在获得已发布的 GO-GPT v2 检查点所作出的预测时，其得分分别为 0.655（强化学习）和 0.622（监督微调），低于仅使用这些预测结果的得分（0.662；这些预测结果来自已发布检查点在 BioReason-Pro 的测试序列副本上运行所得，而在该基准测试自身的测试集上，其得分为 0.666，见表 1）。这与 Fang 等人 [2026] 的发现一致，即该模型的答案遵循其文本输入。RumiFold 在原始 GO-GPT v1 训练集上进行微调后同样表现强劲：在该基准测试的测试集上评估时，其得分达到 0.673，而 GO-GPT v1 为 0.664，BioReason-Pro 为 0.662（E 节）。

它学会了折叠，还是只用上了 ESM-C？ RumiFold 读取 ESM-C 嵌入，而 GO-GPT v1 和 v2 读取 ESM-2（3B 和 650M），因此它的性能提升可能来自更新的编码器，而非学会折叠。三项对照实验对此进行了检验（表 1）。不进行折叠的同一模型。*Qwen + ESM-C* 是基础版 Qwen3-8B，使用相同的桥接模块、LoRA 秩和训练数据，但其适配器从未接受过折叠训练。它的得分为 0.638，而 RumiFold 为 0.683（$p = 10^{-4}$）；在序列一致性低于 50% 时，得分为 0.480，而 RumiFold 为 0.508（$p = 0.0065$）。更新的编码器还不够。我们按照 GO-GPT v2 的原有方案重新训练，仅将编码器从 ESM-2 换成 ESM-C-6B。在已发布检查点对应的轮次（90）时，其得分为 0.651，低于已发布模型的 0.666。用折叠模型的状态代替折叠本身。我们将冻结的 ESMFold2 的主干网络状态提供给采用相同 LoRA 秩的基础版 Qwen3-8B，并使用相同数据对其进行训练。其得分为 0.638，而 RumiFold 为 0.683（$p = 10^{-4}$）；在序列一致性低于 50% 时，差距更大（0.431 对 0.508，$p = 10^{-4}$）。因此，读取专用折叠模型的状态，无法替代让语言模型学会折叠。

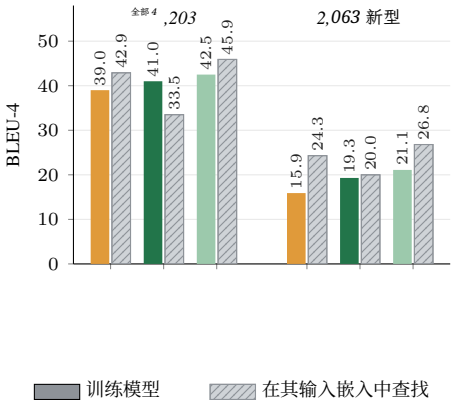
5.2 自由文本功能描述

Prot2Text [Fei 等人， 2025] 要求对蛋白质的功能进行自由文本描述。我们在其训练集上微调 RumiFold，并在其 4,203 个测试蛋白质上进行评估；此外，也单独在 2,063 个新颖 蛋白质上进行评估，这些蛋白质的参考描述未出现在训练集中。我们将其与最强的已发表系统 Prot2Text-V2 进行比较：在相同的蛋白质上使用其已发布的权重，并采用相同的简短提示、贪心解码和评分器。

我们报告三项指标。BLEU-4 衡量生成的一至四词短语相对于参考描述的精确率，并会对短于参考描述的答案施加惩罚。ROUGE-L [Lin， 2004] 是两个序列按顺序共享的最长词序列上的 F 分数。两者都会统计共享词，因此我们还报告 Prot2Text-V2 论文中的评审分数：GPT-4o 读取参考描述和生成的

模型	BLEU-4	R-L	GPT-4o
全部 *4,203* 个测试蛋白质			
SEQ2LLaMA	25.8	35.2	35.2
Prot2Text-V2	39.0	55.7	59.5
RumiFold	41.0	54.5	58.3
RumiFold（L76）	**42.5**	**55.8**	**60.2**
2,063 种新型蛋白质			
SEQ2LLaMA	9.9	16.0	15.6
Prot2Text-V2	15.9	26.4	31.9
RumiFold	19.3	27.0	32.4
RumiFold（L76）	**21.1**	**28.0**	**34.4**

（a）BLEU-4、ROUGE-L 和 GPT-4o评审分数（0–100）。粗体：最佳；下划线：第二佳。RumiFold（L76）使用 ESM-C 的第 76 层，而不是最后一层。



（b）各训练模型的 BLEU-4 分数（实心）与在其读取的嵌入中查找的分数（斜线填充）对比：不经训练，直接复制最近的训练蛋白质的描述。在所有测试蛋白质上，RumiFold 是唯一表现优于查找结果的模型（41.0 对比 33.5）。

图 4：**RumiFold 对新型蛋白质的描述优于 Prot2Text-V2，而读取质量更高的 ESM-C 层可进一步扩大这一优势**。所有模型均使用相同的测试蛋白质、相同的提示（不包含蛋白质名称或分类信息）、贪心解码方式和评分器运行。新型蛋白质是指其参考描述未出现在训练集中的蛋白质。在（b）中，查找采用残基均值和余弦距离；Prot2Text-V2 读取 ESM-2 3B 的最后一层，RumiFold 读取 ESM-C 的最后一层，而 RumiFold（L76）读取其第 76 层。

描述内容，并按 0 到 10 的范围评估预测对功能、具体细节和完整性的捕捉程度；我们报告按 0–100 分制计算的均值。

结果。 与 Prot2Text-V2 相比，RumiFold 对新型蛋白质的描述更出色（图 4a）。在 2,063 个无法从训练数据中回忆出描述的蛋白质上，其 BLEU-4 得分为 19.3，高于对方的 15.9；ROUGE-L 得分（27.0 对 26.4）和评审分数（32.4 对 31.9）则持平或略胜一筹。在所有测试蛋白质上，两者表现相近：RumiFold 的 BLEU-4 得分更高（41.0 对 39.0），但 ROUGE-L 得分（54.5 对 55.7）和评审分数（58.3 对 59.5）较低。RumiFold 在基准测试上的训练量更少，却取得了这些成绩：Prot2Text-V2 经过 12 轮对比对齐训练，随后又进行了 24 轮指令微调。

输入层很重要。 这两个模型读取不同的编码器，而这些编码器在训练开始前所承载的功能信息量也不同。我们通过查找来衡量这一点：对于每个测试蛋白质，我们在模型所读取的嵌入中找到与其最近的训练蛋白质，并将该训练蛋白质的描述赋予它（图 4b）。RumiFold 所读取的 ESM-C 最后一层本身对功能的描述能力较弱：查找得分为 33.5 BLEU-4，而 Prot2Text-V2 所读取的 ESM-2 3B 特征得分为 42.9。在其自身的输入嵌入上，RumiFold 是唯一一个优于查找结果的模型：在所有测试蛋白质上，其得分达到 41.0，而查找结果为 33.5；相比之下，Prot2Text-V2 的生成结果低于其输入的查找结果（39.0 对 42.9）。

读取更好的层。 ESM-C 的较早层承载了更多信息：在第 76 层进行相同的查找，得分为 45.9，高于 ESM-2 3B。因此，我们采用相同方式在第 76 层上对 RumiFold 进行微调，以替代最后一层；这一选择基于验证数据确定，并对其后期检查点的权重取平均。随后，RumiFold（L76）在所有测试蛋白质上、全部三个指标中均领先于 Prot2Text-V2（BLEU-4 为 42.5 对 39.0，ROUGE-L 为 55.8 对 55.7，评审分数为 60.2 对 59.5），不过后两个指标均处于噪声范围内（评审分数差异 $+0.7$，95% 区间 $[-0.1, +1.5]$）。在新型蛋白质上，领先优势很明显：BLEU-4 为 21.1 对 15.9，ROUGE-L 为 28.0 对 26.4，评审分数为 34.4 对 31.9（配对差值 $+2.5$，95% 区间 $[+1.4, +3.6]$）。

关于另一项自由文本描述基准测试（Mol-Instructions）的早期实验见附录 E，其中 RumiFold 在受训练的蛋白质模型中同样处于最先进水平。

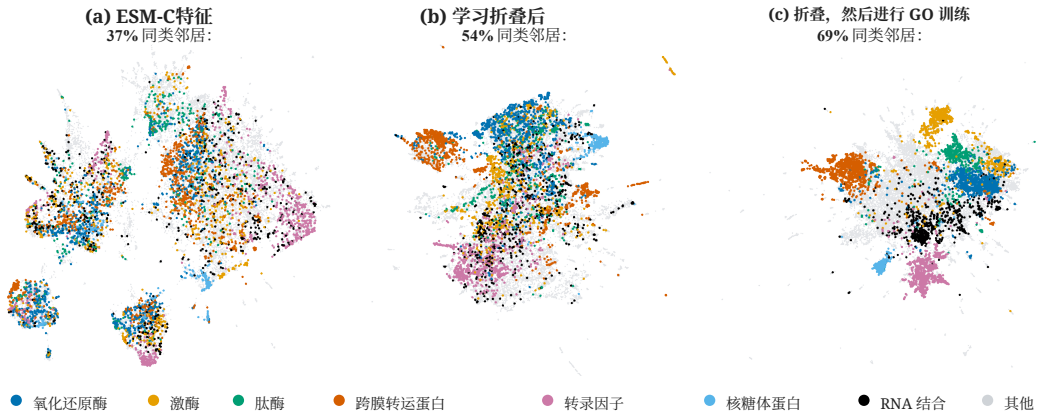


图5：**学习折叠能按功能区分蛋白质，而后续的功能训练会进一步强化这种区分**。对同一批23,084种蛋白质的残基平均表示进行UMAP可视化：**(a)** 模型读取的最终ESM-C层；**(b)** RumiFold-Struct的状态；**(c)** 同一模型经过基因本体微调后的状态。颜色根据真实注释标出七种分子功能类别，每个面板中的颜色一致；不属于任何类别或同时属于多个类别的蛋白质以灰色表示。百分比表示在完整空间（余弦距离）中，660种测试蛋白质的十个最近训练蛋白质里带有其类别的比例；随机情况下为4%。

6 学会折叠除了结构预测还能带来什么?

除了预测结构的能力之外，语言模型还能从学习折叠中获得什么？我们分三个步骤对此展开研究。首先，我们考察折叠如何改变模型对蛋白质结构和功能的内部表征。其次，我们测试语言模型在回答问题时，是否会将这些折叠表示用作内部工作空间。最后，我们探究这一工作空间是否反映了具有生物学意义的机制，将模型的预测与通过实验表征的功能残基联系起来。

6.1 折叠丰富了表征

在使用功能数据进行任何微调之前，折叠就已按功能组织蛋白质。图5在三个表征空间中绘制了同一批蛋白质。在模型接收的ESM-C特征中，各功能类别彼此重叠：蛋白质的最近邻有37%的概率与其属于同一分子功能类别。仅经过折叠训练、从未接触过任何功能标签后，这一比例升至54%，且转运蛋白、转录因子和核糖体蛋白已彼此区分。在折叠训练的基础上进行功能训练后，这一比例达到69%。通过最近邻查找基因本体术语也得出了相同结论（ESM-C为0.435，折叠后为0.455，折叠并进行功能训练后为0.523；见F节）。

残基层面探针。 逐个残基来看，结论也相同。我们分别使用 RumiFold-Struct 和 ESM-C-6B 的冻结特征，在 VenusX 的 17 项残基任务上各拟合一个探针 [Tanet al., 2025], , 并针对测试蛋白质计算配对自助法置信区间（见 F 节的表11）。仅折叠状态在 8 项任务上表现更好，包括所有结构域任务，以及 MF50 和 MP50 划分中的活性位点和保守位点任务；在 2 项任务上表现较差，即跨家族活性位点和保守位点任务。其余 7 项任务的结果没有显著差异。未经折叠、仅针对功能微调的对照模型，其状态在全部 17 项任务上都不如 ESM-C特征，在其中 15 项任务上也不如仅折叠状态。

我们还表明，这种富集是在折叠轮次中计算得出的，并非由编码器传递而来（见 F.2 节）。

6.2 RumiFold 使用其内部折叠工作空间

我们将蛋白质的记忆 定义为第二个折叠过程在每一层为其每个残基计算的键和值（公式（3））：模型折叠时写入的条目，每个残基和每一层各一个。根据设计，语言处理过程只关注这段记忆，不关注蛋白质的其他任何信息。在上一节中，我们展示了折叠过程会围绕结构和功能组织这段记忆

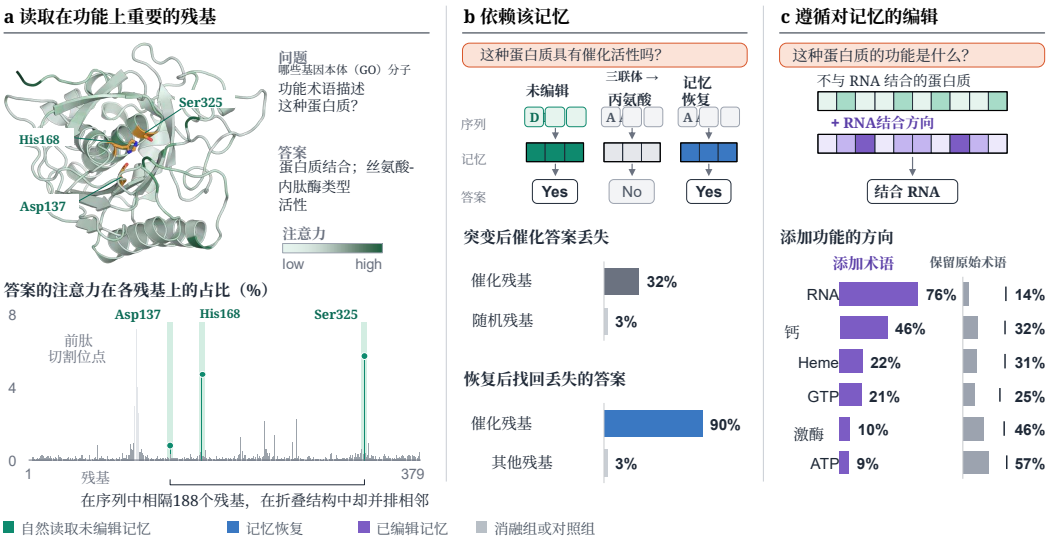


图6：**RumiFold 使用其内部折叠工作空间。（a）**当 RumiFold 说出枯草杆菌蛋白酶的功能时，其注意力集中在催化三联体（Asp137、His168、Ser325）上；这三个残基在序列中相距甚远，却在折叠结构中彼此相邻。**（b）**将酶的催化残基突变为丙氨酸，会使 1,448 个留出酶中 32% 的答案不再是“催化”，而随机残基仅为 3%。仅恢复这些残基处的记忆，就能挽回 90% 的丢失答案；相比之下，在其他残基处恢复记忆只能挽回 3%。**（c）**将某项功能的方向添加到记忆中，会使模型为缺少该功能的蛋白质写出该功能（每项功能约 2,600 个），但代价是模型此前写出的部分术语会减少；刻度线表示大小相同的随机方向。

（6.1节）。在此，我们探究模型如何 使用这段记忆。模型在描述某种功能时，哪些残基会受到最多注意力？恢复这些残基处的记忆，能否找回因突变而改变的答案？编辑值向量中的某个方向，能否改变模型报告的功能？（图6）

答案会读出具有功能重要性的残基。当 RumiFold 回答有关蛋白质功能的问题时，其注意力集中在执行该功能的残基上。在枯草杆菌蛋白酶中，催化三联体在序列中相距甚远，只在折叠结构中相互接近，而答案捕捉到了这一点：在 379 个残基中，Ser325 和 His168 获得的注意力份额分别位列第二和第三，Asp137 位列第十七；全部注意力的 31% 落在三联体周围 10 埃范围内，而这些区域仅涉及 17% 的残基（图6a）。获得最多注意力的单个残基同样具有功能：它是前肽的最后一个残基，酶在此处切割自身。这一模式具有普遍性。在 369 个具有诱变数据的留出酶中，使活性完全消失的突变残基在其蛋白质内部的注意力排名处于第 89 百分位，使活性降低的突变残基处于第 81 百分位，而没有报告影响的突变残基处于第 70 百分位。

答案取决于这种记忆。模型会高度关注活性位点和催化残基，但它的答案真的取决于这些残基吗？为检验这一点，我们将一种酶中已注释的催化残基及其两侧各两个相邻残基突变为丙氨酸，然后让模型从头训练，对突变体进行编码和折叠。在 1,448 种未编辑时被判定具有催化功能的留出酶中，32% 不再得到这一答案；突变相同数量的其他残基时，这一比例为 3%（图 6b）。直接测量概率时，突变使 log P(Yes) 平均降低 0.62，而对照组仅降低 0.03。这个问题有意设置得较为宽泛，而且突变后的酶仍保有酶的折叠和家族特征，因此大多数答案理应不变。然而，在保留突变序列、仅恢复催化残基处原始记忆的情况下，答案又回来了：这种替换挽回了突变所导致答案变化的 90%（表12），使 96% 的答案重新变为肯定，且 log P(Yes) 与未编辑时的值相差不到 0.09。在其他位置恢复相同数量的条目，只能挽回 3%，仍只有 68% 的答案为肯定，并不比突变体更好。记忆中位于负责化学反应的残基处的少数条目，就足以承载这一答案。G节列出了所有实验条件。

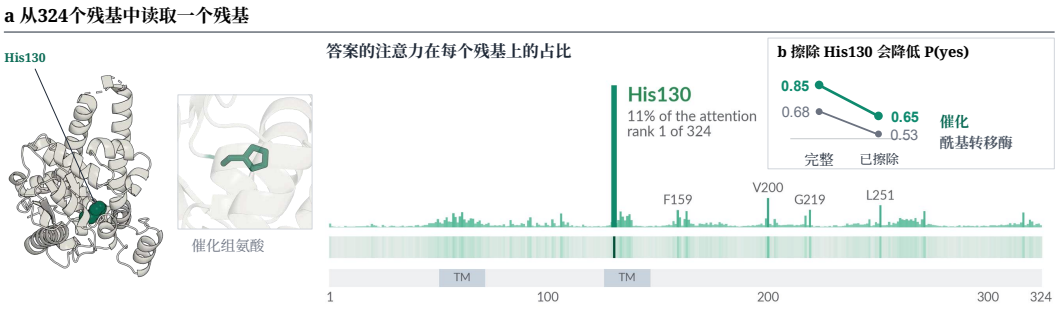


图 7：**RumiFold 读取了 TMEM68 的活性位点；该蛋白未纳入训练。(a)** 读取了 *324* 个残基中的一个。当 RumiFold 判断人类 TMEM68 的功能时，答案的注意力有 11% 指向 His130，高于其余 323 个残基中的任何一个。左：预测结构（AlphaFold 数据库）中的 His130（催化组氨酸），附局部放大图。右：沿序列分布的注意力占比；TM 表示跨膜片段。**(b)** 擦除 *His130* 会降低答案的肯定概率。从折叠记忆中移除 His130，会使回答“是”的概率从 0.85 降至 0.65（催化活性），并从 0.68 降至 0.53（酰基转移酶活性）。序列一致性达到或超过 30% 的 TMEM68 及其同源蛋白，均未纳入残基和功能训练；鉴定出 His130 的实验发表于 2023 [Wang 等人，2023a，McLelland 等人，2023]。

答案会随对记忆的编辑而改变。如果答案是从记忆中读取的，那么写入记忆就应该能改变答案。对于五种分子功能中的每一种，我们在记忆的值空间中取一个方向：具有该功能的蛋白质均值减去不具有该功能的蛋白质均值；然后将这一方向加到每个符合条件且缺少该功能的留出蛋白质的记忆中，每种功能约涉及 2,600 个蛋白质。当被问及蛋白质的功能时，模型随后会写入添加的功能（图 6c）：对于加入 RNA结合方向的蛋白质，这一比例为 76%；对于钙结合，为 46%；对于 ATP结合，则降至 9%，是五种功能中最低的。相同大小的随机方向从未添加过该术语。这种编辑会改变答案，但并不精确：它也会扰乱其他答案。编辑后，蛋白质仅保留编辑前所写术语的 14% 至 57%；而采用随机方向时，这一比例为 84% 至 89%。

这些结果共同表明，折叠记忆是语言模型据以进行推理的工作空间。答案会读取关键残基；当这些残基的记忆有误时，答案会失败；恢复这些记忆后，答案也会恢复；重写记忆后，答案则会发生变化。G 和 G.1 节给出了所有条件和对照，包括负责读取的注意力头。

6.3 工作空间与生物学机制相一致

如果答案建立在承载某项功能的残基之上，那么读取模型就能够用来询问某项功能存在于蛋白质的何处。我们沿着 6.2 节的三个步骤考察一种蛋白质之所以选择它，是因为它及其同源蛋白被排除在训练集之外，而且其催化残基直到 2023年才通过实验得到鉴定。

一种后来才发现其功能的蛋白质。 人类 TMEM68 已知属于甘油磷脂酰基转移酶家族，但其酶学功能一直未被确定，直到 2023年，研究才表明它能够合成三酰甘油，并且当单个组氨酸 His130 发生突变时会失去该活性 [Wang 等，2023a，McLelland 等人， 2023]。TMEM68 属于 GO 基准中的 CAFA 式时间留出集：在 2022年 11 月之前，它没有实验功能注释，并于 2023年获得了一项注释。它及其所有序列一致性为 ≥30% 的同源蛋白也被排除在我们的残基和功能训练之外。其他训练蛋白质的注释快照来自 2026 年，因此这是蛋白质层面的留出，而不是全局知识截止点。基因本体微调集确实包含其牛同源直系同源蛋白（93% 一致），但该蛋白仅带有通用膜相关术语，没有功能或位点标签。

该模型读取 His130，其答案取决于该残基。 当用通俗语言询问 TMEM68 是否具有催化活性时，RumiFold 以 0.85 的概率回答“是”；当要求其说出

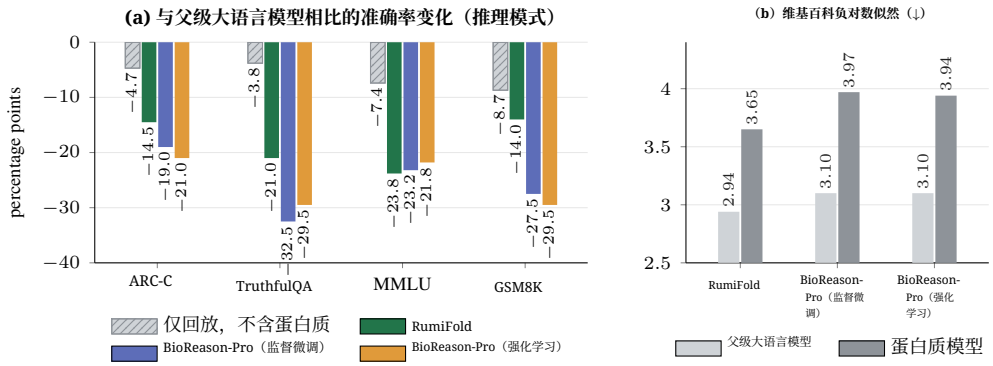


图8：**蛋白质训练对通用语言能力的影响。(a)** 在包含逐步推理的情况下，相对于每个模型自身父级大语言模型的准确率变化，基于包含200个项目的子集（区间约为 ±6 个百分点）。**(b)** 在1,578段 WikiText 文本（约259k个词元）上的每词元负对数似然（纳特）；每个模型都在其自身聊天模板内阅读文本。两个面板中的 RumiFold 均为训练95k步的检查点。(a)中的阴影柱：仅使用回放目标训练的基础模型，未使用蛋白质数据，训练约44k步（在完整测试集上进行少样本直接作答，因此仅用于与其他柱状图进行尺度比较）。

功能时，它写下了酰基转移酶活性。His130 获得了答案的注意力的 11%，超过该蛋白质 324 个残基中的任何其他残基，约为排名第二者所占比例的五倍（图7a）；在小鼠直系同源蛋白中，同一残基也排名第一。从折叠记忆中擦除这一个残基后，概率降至 0.65；更具体的答案“酰基转移酶活性”则从 0.68 降至 0.53（图7b）。擦除 24 个对照残基中的任意一个——每个残基距离 His130 至少 12埃，且埋藏程度相同——两个答案均未发生超出读出分辨率范围的变化。

该残基正是实验发现的那个残基。 模型没有获得该蛋白质的任何位点注释，但它在 TMEM68 的14个组氨酸中选定了一个残基，其中7个位于其酰基转移酶结构域中。该残基正是其突变会在实验室中消除活性的那个残基[Wang 等人，2023a，McLelland 等人， 2023]。His130 也是通过将 TMEM68 与相关酶进行比较而预期得到的残基，因此这并不是一次发现。它所展示的是：在训练中没有获得该蛋白质的位点标签的情况下，模型的内部工作空间仍然与执行该化学反应的残基相一致。

这是一个选定蛋白质的案例，我们并不声称模型的注意力通常能找到活性位点。在 369 种留出酶中，抹去已注释的活性位点残基，对同一答案的影响远小于抹去模型最关注的残基；而注意力区分注释位点与其他残基的能力，只与埋藏程度相当。这个案例表明，可以逐个残基地探查工作空间，而且探查结果确实可能触及真实的生物化学。G.4节提供了更多案例，其中包括读出未能命中的案例，以及一个未经检验的假设。

学会折叠带来的能力。学会折叠赋予语言模型三项超越结构本身的能力。折叠丰富了它的表征：在接受任何功能训练之前，蛋白质和残基就已按功能组织起来（6.1节）。模型会利用折叠留下的工作空间：它的答案会读取关键残基，取决于其中存储的信息，并随之进行的编辑而变化（6.2节）。而且，这个工作空间与生物学机制相吻合：在一种未参与训练的蛋白质中，答案所依据的残基正是实验表明具有催化作用的残基（6.3节）。G节列出了突变实验的所有对照，G.1节则追踪了记忆中的单一功能。

7 局限性

输入一层。 RumiFold 的输入仅为 ESM-C 的最后一层。我们从一开始就选择了它，因为这是标准输入，也是 GO-GPT 和 Prot2Text-V2 从其编码器中获取的输入。对编码器的各层取均值或采用可学习的 softmax 混合，会是更好的输入：在我们测试的用于检索的表示中，最后一层效果最差（第 5 节），而中间层包含更多功能

信息（F 节）。我们没有足够的存储空间或算力来保留折叠语料库中每一层的特征，因此将此留待未来工作。

代价是什么。 教会语言模型利用自身权重对蛋白质进行计算并非没有代价。尽管在整个训练过程中回放普通对话，RumiFold 在 MMLU [Hendrycks 等人、 2020], ARC-Challenge [Clark 等人、 2018], TruthfulQA[Lin 等人、 2021] 和 GSM8K [Cobbe 等人这四项基准测试上的得分，仍比基础模型低 14 到 21 分 2021], ，降幅与 BioReason-Pro 相对于其自身父模型的降幅相当（图8）。部分代价来自回放本身：仅针对回放目标进行训练、未使用蛋白质数据的同一批适配器，在这四项基准测试上的得分也低了 4 到 9 分。表 2 列出了所有数据。

隐藏状态是否必不可少? 我们主张在模型的隐藏状态中加工蛋白质。如今，另一条路径也逐渐显现：近期研究表明，仅通过文本就能编辑模型的权重 [Takehi 等人， 2026], 这意味着，我们在此通过隐藏状态表征所获得的一些成果，或许也能通过文本实现。我们尚未比较这两种方法，并认为它们是互补的研究方向。

固定的轮次数量。RumiFold 始终分两轮进行折叠。允许轮次数量变化，并循环执行“折叠、用文本推理、再折叠”，会更合理。将轮次数量固定为两轮，也使第一轮的作用有限。

8 讨论

我们的研究旨在探究通用语言模型能否学会折叠蛋白质，以及如果能够，它能从中获得什么。
RumiFold 表明这是可行的：一个冻结的 Qwen3-8B，搭配低秩适配器和线性输出头，其蛋白质折叠效果与专用单表征模型大致相当。折叠的用途不止于结构：经过折叠预训练后，RumiFold 在基因本体蛋白质功能预测和自由文本功能描述任务上取得了最先进的结果，领先于从未学过折叠的匹配模型。而且，该模型会利用折叠所构建的内容：它的答案会读取折叠轮次留下的逐残基记忆，依赖于这些记忆，并会随着记忆的编辑而改变。
RumiFold 提出了一种构建科学语言模型的更广泛方法：将基础科学任务用作预训练，以发展支持下游推理的内部表征。折叠只是众多计算中的一种。进化推断、设计和结合也可以用同样的方式学习，并形成各自的工作空间。我们还设想语言与生物分子计算交替进行的模型。这样的模型可以在自身的推理过程中运行实验：提出一个突变，对其进行折叠，读取发生了哪些变化，然后决定下一步尝试什么。它的思维链将包含结构，也包含句子。更困难的问题会获得更多这类内部实验，因此，更长时间的思考意味着进行更多计算，而不仅仅是写下更多内容。

我们将 RumiFold 视为迈向原生生物分子智能的第一步：在这类语言模型中，生物分子计算与语言推理并非通过界面连接起来的独立系统，而是同一个模型协同运作的能力。

致谢与资金披露

本材料基于美国国家科学基金会研究生研究奖学金计划的资助，资助编号 DGE-2439018。本材料中表达的任何观点、研究发现、结论或建议均属于作者，不一定反映美国国家科学基金会的观点。感谢 Goodfire 通过 Goodfire 研究资助支持这项工作，并提供对其可解释性研究智能体 Silico 的使用权限。本工作还得到了阿尔弗雷德·P·斯隆基金会的支持。感谢 Bau 与 Jin 实验室成员提出的反馈，以及 Sheng-Yu Wang 对图表提出的宝贵建议。

参考文献

Josh Abramson、Jonas Adler、Jack Dunger、Richard Evans、Tim Green、Alexander Pritzel、Olaf Ronneberger、Lindsay Willmore、Andrew J. Ballard、Joshua Bambrick 等。利用 AlphaFold 3 准确预测生物分子相互作用的结构。《自然》，630(8016):493–500，2024年。doi：10.1038/s41586-024-07487-w。

Michael Ashburner、Catherine A. Ball、Judith A. Blake、David Botstein、Heather Butler、J. Michael Cherry、Allan P. Davis、Kara Dolinski、Selina S. Dwight、Janan T. Eppig、Midori A. Harris、David P. Hill、Laurie Issel-Tarver、Andrew Kasarskis、Suzanna Lewis、John C. Matese、Joel E. Richardson、Martin Ringwald、Gerald M. Rubin 和 Gavin Sherlock。基因本体：统一生物学的工具。《自然·遗传学》，25(1):25–29，2000年5月。ISSN 1546-1718。doi：10.1038/75556。URL http://dx.doi.org/10.1038/75556。

Sankar Basu 和 Björn Wallner。《DockQ：蛋白质-蛋白质对接模型的质量衡量指标》。《*PLOS ONE*》，11(8)：e0161879，2016年8月。ISSN 1932-6203。doi：10.1371/journal.pone.0161879。URL：http://dx.doi.org/10.1371/journal.pone.0161879。

Peter Clark、Isaac Cowhey、Oren Etzioni、Tushar Khot、Ashish Sabharwal、Carissa Schoenick 和 Oyvind Tafjord。你以为自己解决了问答问题？试试 ARC，即 AI2 推理挑战。*arXiv* 预印本 *arXiv:1803.05457*，2018年。

Karl Cobbe、Vineet Kosaraju、Mohammad Bavarian、Mark Chen、Heewoo Jun、Lukasz Kaiser、Matthias Plappert、Jerry Tworek、Jacob Hilton、Reiichiro Nakano、Christopher Hesse 和 John Schulman。训练验证器来解决数学文字题。*arXiv* 预印本 *arXiv:2110.14168*，2021年。

ESM 团队。ESM Cambrian：通过无监督学习揭示蛋白质的奥秘。EvolutionaryScale，2024年。https://www.evolutionaryscale.ai/blog/esm-cambrian。

Adibvafa Fallahpour、Arman Seyed-Ahmadi、Parsa Idehpour、Omar Ibrahim、Benedict M.H. Choi、Purav Gupta、Jack Naimer、Kevin Zhu、Abhinav Adduri、Arnav Shah、Shihao Ma、Talu Güloglu、Nuo Liu、Haotian Cui、Arihant Jain、Max de Castro、Amirfaham Fallahpour、Antonio Cembellin-Prieto、John S. Stiles、Filip Nemˇcko、Alexander A. Nevue、Hyungseok C. Moon、Lucas Sosnick、Olivia Markham、Haonan Duan、Michelle Y. Y. Lee、Andrea F. M. Salvador、Chris J. Maddison、Christoph A. Thaiss、Chiara Ricci-Tam、Brian S. Plosky、Dave P. Burke、Patrick D. Hsu、Hani Goodarzi 和 Bo Wang。BioReason-Pro：通过多模态生物学推理推进蛋白质功能预测。*bioRxiv*，2026 年3月。doi: 10.64898/2026.03.19.712954。URL http: //dx.doi.org/10.64898/2026.03.19.712954。

Ada Fang、Nikitha Thoduguli、Lukas Fesser、Hanlin Zhang、ShamM. Kakade 和 Marinka Zitnik。生物学推理模型何时会使用其生物学输入？*arXiv* 预印本 *arXiv:2610.00898*，2026 年。

Yin Fang、Xiaozhuan Liang、Ningyu Zhang、Kangwei Liu、Rui Huang、Zhuo Chen、Xiaohui Fan 和 Huajun Chen。Mol-Instructions：面向大语言模型的大规模生物分子指令数据集。*arXiv* 预印本 *arXiv:2306.08018*，2023年。

Xiao Fei、Michail Chatzianastasis、Sarah Almeida Carneiro、Hadi Abdine、Lawrence P. Petalidis 和 Michalis Vazirgiannis。《Prot2Text-V2：基于多模态对比对齐的蛋白质功能预测》。*arXiv* 预印本 *arXiv:2505.11194*，2025。

Leo Gao、Tom Dupré la Tour、Henk Tillman、Gabriel Goh、Rajan Troll、Alec Radford、Ilya Sutskever、Jan Leike 和 Jeffrey Wu。《稀疏自编码器的扩展与评估》。*arXiv* 预印本 *arXiv:2406.04093*，2024年。

A. Ghafarollahi 和 M. J. Buehler。《ProtAgents：通过结合物理学与机器学习的大语言模型多智能体协作实现蛋白质发现》。*arXiv* 预印本 *arXiv:2402.04268*，2024年。

Thomas Hayes、Roshan Rao、Halil Akin、Nicholas J. Sofroniew、Deniz Oktay、Zeming Lin、Robert Verkuil、Vincent Q. Tran、Jonathan Deaton、Marius Wiggert、Rohil Badkundri、Irhum Shafkat、Jun Gong、Alexander Derry、Raul S. Molina、Neil Thomas、Yousuf A. Khan、Chetan Mishra、Carolyn Kim、Liam J. Bartie、Matthew Nemeth、Patrick D. Hsu、Tom Sercu、Salvatore Candido 和 Alexander Rives。借助语言模型模拟 5 亿年的进化。《科学》，387(6736):850–858，2025年2月。ISSN 1095-9203。doi: 10.1126/science.ads0018。URL http://dx.doi.org/10.1126/science.ads0018。

Dan Hendrycks、Collin Burns、Steven Basart、Andy Zou、MantasMazeika、DawnSong 和 Jacob Steinhardt。《大规模多任务语言理解评测》。*arXiv* 预印本 *arXiv:2009.03300*，2020年。

Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、SheanWang、Lu Wang 和 Weizhu Chen。《LoRA：大语言模型的低秩适配》。*arXiv* 预印本 *arXiv:2106.09685*，2021年。

Kexin Huang、Serena Zhang、Hanchen Wang、Yuanhao Qu、Yingzhou Lu、Yusuf Roohani、Ryan Li、Lin Qiu、Gavin Li、Junze Zhang、Di Yin、Shruti Marwaha、Jennefer N. Carter、Xin Zhou、Matthew Wheeler、Jonathan A. Bernstein、Mengdi Wang、Peng He、Jingtian Zhou、Michael Snyder、Le Cong、Aviv Regev 和 Jure Leskovec。《Biomni：通用生物医学人工智能智能体》。*bioRxiv*，2025年6月。doi：10.1101/2025.05.30.656746。URL：http://dx.doi.org/10.1101/2025.05.30. 656746。

John Jumper、Richard Evans、Alexander Pritzel、Tim Green、Michael Figurnov、Olaf Ronneberger、Kathryn Tunyasuvunakool、Russ Bates、Augustin Žídek、Anna Potapenko、Alex Bridgland、Cle mens Meyer、Simon A. A. Kohl、Andrew J. Ballard、Andrew Cowie、Bernardino Romera- Paredes、Stanislav Nikolov、Rishub Jain、Jonas Adler、Trevor Back、Stig Petersen、David Reiman、Ellen Clancy、Michal Zielinski、Martin Steinegger、Michalina Pacholska、Tamas Berghammer、Sebastian Bodenstein、David Silver、Oriol Vinyals、Andrew W. Senior、Koray Kavukcuoglu、Push- meet Kohli 和 Demis Hassabis。《利用 AlphaFold 高精度预测蛋白质结构》《自然》，596(7873):583–589，2021 年 7 月。ISSN 1476-4687。doi: 10.1038/s41586-021-03819-2。URL http://dx.doi.org/10.1038/s41586-021-03819-2。

Andriy Kryshtafovych、Torsten Schwede、Maya Topf、Krzysztof Fidelis 和 John Moult。蛋白质结构预测方法的关键评估（CASP）——第十四轮。《蛋白质：结构、功能与生物信息学》，89(12):1607–1617，2021年10月。ISSN 1097-0134。doi: 10.1002/ prot.26237。URL http://dx.doi.org/10.1002/prot.26237。

Andriy Kryshtafovych、Torsten Schwede、Maya Topf、Krzysztof Fidelis 和 John Moult。计算结构生物学中深度学习的进展与瓶颈：CASP 第十六轮。《蛋白质：结构、功能与生物信息学》，94(1):5–14，2025年11月。ISSN 1097-0134。doi: 10.1002/prot.70076。URL http://dx.doi.org/10.1002/prot.70076。

Nathan Lambert、Jacob Morrison、Valentina Pyatkin、Shengyi Huang、Hamish Ivison、FaezeBrahman、Lester James V. Miranda、Alisa Liu、Nouha Dziri、Shane Lyu、Yuling Gu、Saumya Malik、Victoria Graf、Jena D. Hwang、Jiangjiang Yang、Ronan Le Bras、Oyvind Tafjord、Chris Wilhelm、Luca Soldaini、Noah A. Smith、Yizhong Wang、Pradeep Dasigi 和 Hannaneh Hajishirzi。《Tulu 3：推动开放式语言模型后训练的前沿》。*arXiv* 预印本 *arXiv:2411.15124*，2024年。

Chin-Yew Lin。《ROUGE：用于摘要自动评估的软件包》。载于文本摘要的拓展，第 74–81 页。计算语言学协会，2004 年。

Stephanie Lin、Jacob Hilton 和 Owain Evans。《TruthfulQA：衡量模型模仿人类谬误的程度》。*arXiv* 预印本 *arXiv:2109.07958*，2021年。

Zeming Lin、Halil Akin、Roshan Rao、Brian Hie、Zhongkai Zhu、Wenting Lu、Nikita Smetanin、Robert Verkuil、Ori Kabeli、Yaniv Shmueli、Allan dos Santos Costa、Maryam Fazel-Zarandi、Tom Sercu、Salvatore Candido 和 Alexander Rives。使用语言模型进行进化尺度的原子级蛋白质结构预测。《科学》，379(6637):1123–1130，2023年3月。ISSN 1095-9203。doi：10.1126/science.ade2574。URL http://dx.doi.org/10.1126/science.ade2574。

Nuowei Liu、Changzhi Sun、Tao Ji、Junfeng Tian、Jianxin Tang、Yuanbin Wu 和 Man Lan。《EvoL- lama：通过多模态结构和序列表示增强大语言模型对蛋白质的理解》。*arXiv* 预印本 *arXiv:2412.11618*，2024年。

Yizhen Luo、Jiahuan Zhang、Siqi Fan、Kai Yang、Yushuai Wu、Mu Qiao 和 Zaiqing Nie。《BioMedGPT：面向生物医学的开放式多模态生成式预训练变换器》。*arXiv* 预印本*arXiv:2308.09442*，2023年。

Gian-Luca McLelland、Marta Lopez-Osias、Cristy R. C. Verzijl、Brecht D. Ellenbroek、Rafaela A. Oliveira、Nicolaas J. Boon、Marleen Dekker、Lisa G. van den Hengel、Rahmen Ali、Hans Janssen、Ji-Ying Song、Paul Krimpenfort、Tim van Zutphen、Johan W.Jonker 和 ThijnR. Brummelkamp。《一种替代性甘油三酯生物合成途径的发现》。《自然》，621(7977):171–178，2023年。doi: 10.1038/s41586-023-06497-4。

Roshan Rao、Joshua Meier、Tom Sercu、Sergey Ovchinnikov 和 Alexander Rives。Transformer 蛋白质语言模型是无监督的结构学习器。*bioRxiv*，2020年12月。doi：10.1101/2020.12.15.422761。URL：http://dx.doi.org/10.1101/2020.12.15.422761。

Xavier Robin、Juergen Haas、Rafal Gumienny、Anna Smolinski、Gerardo Tauriello 和 Torsten Schwede。持续自动化模型评估 (CAMEO)——关于全自动结构预测方法评估未来的展望。《蛋白质：结构、功能与生物信息学》，89(12):1977–1986，2021年8月。ISSN 1097-0134。doi：10.1002/prot.26213。URL：http://dx.doi.org/10.1002/prot.26213。

Yiqing Shen、Zan Chen、Michail Mamalakis、Yungeng Liu、Tianbin Li、Yanzhou Su、Junjun He、Pietro Liò 和 Yu Guang Wang。TourSynbio：一种多模态大模型与智能体框架，用于桥接文本与蛋白质序列以实现蛋白质工程。*arXiv* 预印本 *arXiv:2408.15299*，2024年。

Kyle Swanson、Wesley Wu、Nash L. Bulaong、John E. Pak 和 James Zou。人工智能智能体虚拟实验室设计出新型 SARS-CoV-2 纳米抗体。《自然》，646(8085):716–723，2025年7月。ISSN 1476-4687。doi: 10.1038/s41586-025-09442-9。URL http://dx.doi.org/10.1038/ s41586-025-09442-9。

Rikiya Takehi、Jayden Teoh、Alex Zhang 和 Omar Khattab。权重空间编程的初步探索：大语言模型能够原生读取和写入神经网络权重。博客文章，2026年11月。https: // rikiyat.github.io/blog/weight-space-programming/。

Yang Tan、Wenrui Gou、Bozitao Zhong、Liang Hong、Huiqun Yu 和 Bingxin Zhou。VenusX：解锁蛋白质精细功能理解。*arXiv* 预印本 *arXiv:2505.11812*，2025。

Xinyou Wang、Zaixiang Zheng、Fei Ye、Dongyu Xue、Shujian Huang 和 Quanquan Gu。DPLM-2：多模态扩散蛋白质语言模型。*arXiv* 预印本 *arXiv:2410.13782*，2024年。

Yu Wang、Fansi Zeng、Zheng Zhao、Lin He、Xiaohong He、Huimin Pang、Feifei Huang 和 Pingan Chang。跨膜蛋白68作为单酰甘油酰基转移酶和二酰甘油酰基转移酶酶参与三酰甘油生物合成。《国际分子科学杂志》，24(3):2012，2023a。doi: 10.3390/ ijms24032012。

Yuyang Wang、Jiarui Lu、Navdeep Jaitly、Josh Susskind 和 Miguel Angel Bautista。SimpleFold：蛋白质折叠比你想象的更简单。*arXiv* 预印本 *arXiv:2509.18480*，2025a。

Zeyuan Wang、Qiang Zhang、Keyan Ding、Ming Qin、Xiang Zhuang、Xiaotong Li 和 Huajun Chen。InstructProtein：通过知识指令对齐人类语言与蛋白质语言。*arXiv* 预印本 *arXiv:2310.03269*，2023b。

Zhicong Wang、Zicheng Ma、Ziqiang Cao、Changlong Zhou、Jun Zhang 和 Yiqin Gao。Prot2Chat：通过文本、序列和结构早期融合构建的蛋白质大语言模型。*arXiv* 预印本 *arXiv:2502.06846*，2025b。

Juntong Wu, Zijing Liu, He Cao, Li Hao, Bin Feng, Zishan Shu, Ke Yu, Li Yuan 和 Yu Li。《重新思考基于文本的蛋白质理解：检索还是大语言模型？》载于*2025*年自然语言处理实证方法会议论文集，第 23737–23757 页。计算语言学协会，2025。doi: 10.18653/v1/2025.emnlp-main.1211。URL：http://dx.doi.org/10.18653/v1/2025.emnlp-main.1211。

Sheng Xu、Qiantai Feng、Lifeng Qiao、Hao Wu、Tao Shen、Yu Cheng、Shuangjia Zheng 和 Siqi Sun。使用 FoldBench 对全原子生物分子结构预测进行基准测试。《自然·通讯》，17(1):442，2025年12月。ISSN 2041-1723。doi: 10.1038/s41467-025-67127-3。URL：http://dx.doi.org/10.1038/s41467-025-67127-3。

An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv、Chujie Zheng、Dayiheng Liu、Fan Zhou、Fei Huang、Feng Hu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang、Jian Yang、Jianhong Tu、Jianwei Zhang、Jianxin Yang、Jiaxi Yang、Jing Zhou、Jingren Zhou、Junyang Lin、Kai Dang、Keqin Bao、Kexin Yang、Le Yu、Lianghao Deng、Mei Li、Mingfeng Xue、Mingze Li、Pei Zhang、Peng Wang、Qin Zhu、Rui Men、Ruize Gao、Shixuan Liu、Shuang Luo、Tianhao Li、Tianyi Tang、Wenbiao Yin、Xingzhang Ren、Xinyu Wang、Xinyu Zhang、Xuancheng Ren、Yang Fan、Yang Su、Yichang Zhang、Yinger Zhang、Yu Wan、Yuqiong Liu、Zekun Wang、Zeyu Cui、Zhenru Zhang、Zhipeng Zhou 和 Zihan Qiu。《Qwen3技术报告》。*arXiv* 预印本 *arXiv:2505.09388*，2025 年。

Naihui Zhou、Yuxiang Jiang、Timothy R. Bergquist 等。《CAFA挑战赛》报告称，通过实验筛选改进了蛋白质功能预测，并为数百个基因提供了新的功能注释。《基因组生物学》，20(1):244，2019年11月。ISSN 1474-760X。doi: 10.1186/s13059-019-1835-8。URL http://dx.doi.org/10.1186/s13059-019-1835-8。

Xibin Zhou、Chenchen Han、Yingqi Zhang、Huan Du、Jiayuan Tian、Jin Su、Renju Liu、Kai Zhuang、Shiyu Jiang、Anthony Gitter、Li Liu、Huayu Li、Meiqi Wu、Shiyang You、Zichen Yuan、Feng Ju、Huilin Zhang、Wei Zheng、Fengyuan Dai、Yuyang Zhou、Yuyang Tao、Dan Wu、Zongze Shao、Yang Liu、Hongyuan Lu 和 Fajie Yuan。使用 Evolla 解码蛋白质的分子语言。*bioRxiv*，2025年1月。doi: 10.1101/2025.01.05.630192。URL http://dx.doi.org/10.1101/ 2025.01.05.630192。

A 架构和训练细节

桥接模块。 公式（1）省略了一项。完整形式中，主干网络的输入为

$$\mathbf{h}_i^0 = \lambda\Big(\mathrm{LayerNorm}\big(W\,\mathrm{LayerNorm}(\mathbf{e}_i) + \mathbf{b}\big) + \mathbb{1}[\,i \text{ starts a chain}\,]\,\mathbf{c}\Big), \tag{4}$$

其中，$\mathbf{c}$ 是一个学习向量，用于标记每条链的第一个残基。

回收。 第一次折叠轮次在不计算梯度的情况下运行：在公式（2）中，$\mathbf{s}_i^{(1)}$ 被视为常量，因此折叠损失仅通过第二次折叠轮次训练主干网络。模型进行一次回收。

位置。 旋转位置索引遵循链内的残基编号，并在链与链之间跳跃 200；在两次折叠轮次和问题轮次中均使用相同的索引。文本位置紧接在最后一个残基之后继续编号。

问题轮次中的上下文布局。 Qwen 的聊天模板的开头词元占据位置 0，位于蛋白质之前；模板的其余部分（一个禁用思考、无系统提示词的单一用户轮次）位于蛋白质之后。文本词元既关注这个开头词元，也关注各个残基。

缓存的键和值。 公式（3）中的蛋白质条目由主干网络自身的第-ℓ 层模块（注意力前归一化、带有 LoRA适配器的键和值投影，以及旋转编码）生成，这些模块应用于第二次折叠过程中块 ℓ 的输入，涵盖全部 36 个块。不会向其中添加任何内容。文本损失的梯度会通过缓存流入第二次折叠过程。

残基指针。 当问题通过地址（例如“A/129”）指定某个残基时，该残基的向量会作为额外的输入嵌入，直接插入地址之后。该向量是将公式（2）应用于第二次折叠过程中最终的隐藏状态，而不是第一次折叠过程中的最终隐藏状态所得。函数问题不包含地址，因此不受影响。

A.1 注意力掩码

图9展示了第3.1节三次传递中使用的两种注意力掩码。

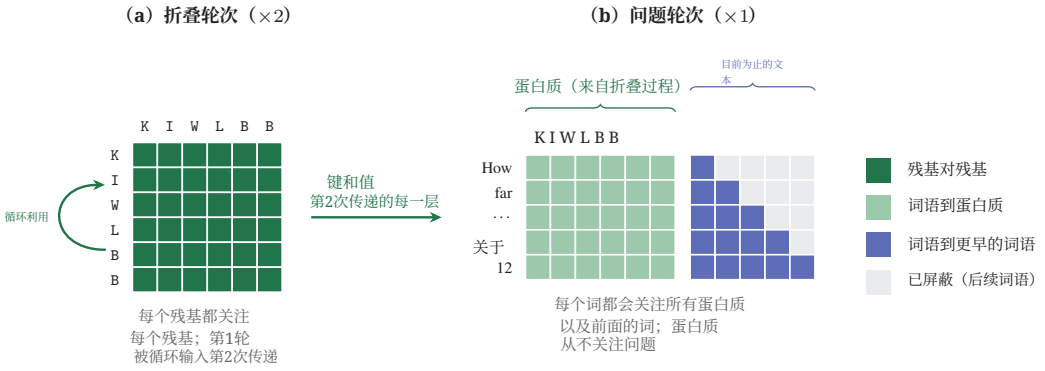


图9：**一个主干网络，三次传递，两个注意力掩码。** 行表示执行注意力的词元，列表示被关注的词元。**(a)** 两次折叠过程仅处理蛋白质，并采用完全双向注意力；第一次的结果会传递给第二次。**(b)** 问题轮次以因果方式处理文本，而且每个词语还会在每一层关注第二次折叠过程为每个残基计算的键和值。

A.2 折叠损失与优化

折叠损失。 无论预测结构经过怎样的旋转或空间平移，都应视为正确，因此我们依次从每个残基的视角比较结构。以残基 i 的参考系为基准，我们询问其他每个残基 j 的每个原子在预测结构和真实结构（标记为 *）中的位置，并测量两者之间的差异：

$$d_{ij,a} = \left\| T_i^{-1}\mathbf{p}_{j,a} - T_i^{\star\,-1}\mathbf{p}^{\star}_{j,a} \right\|, \qquad \mathcal{L}_{\text{frame}} = \underset{i,\,j,\,a}{\text{mean}}\ \min(d_{ij,a},\ 10\,\text{Å}). \tag{5}$$

这是 AlphaFold2 [Jumper 等人， 2021]引入的帧对齐误差；将其上限设为 10 埃，可避免少数位置严重错误的残基占据主导地位。[2] 我们分别计算链内残基对和不同链之间残基对的误差，从而避免复合物的界面被数量多得多的链内残基对淹没。两个小型项用于保持链的物理合理性：$\mathcal{L}_{\text{bond}}$ 惩罚连续残基之间长度偏离真实值的键，以及$\mathcal{L}_{\text{clash}}$ 惩罚距离小于 1.5 埃的原子。完整目标函数为

$$\mathcal{L}_{\text{fold}} = \mathcal{L}^{\text{within}}_{\text{frame}} + \mathcal{L}^{\text{between}}_{\text{frame}} + 0.5\,\mathcal{L}_{\text{bond}} + 0.01\,\mathcal{L}_{\text{clash}}. \tag{6}$$

优化。 AdamW，学习率为 10^{-4}，权重衰减为 0.01，有效批次大小为 24 个组装体，在 8 台 A100 GPU 上进行训练，梯度裁剪阈值为 1.0。不使用权重平均。

A.3 训练数据和留出集

折叠。 折叠数据详见第 3.2 节。我们使用截至 2022年 发布的所有 PDB 条目。对于 AFDB 训练语料库，我们移除与 CAMEO22 或 CASP14靶标 的序列一致性至少为 30% 的每条链及其所在的簇。

[2]实际上，十分之一的　未封顶误差会被加回来，以便远距离误差仍能传　　播梯度。

功能训练。 第 3 阶段数据池始于五个蛋白质–文本来源（Prot-Inst-OOD、Prot2Text、UniProtQA、BioReason-Pro CAFA 和 GO-GPT）的训练划分的并集，共包含 638,317 条不同序列。我们将其与 Prot2Text 的测试集（4,203 个蛋白质）、BioReason-Pro CAFA 基准测试（8,528 个）、我们自己的结构评估集（2,764 个）、VenusX 的验证和测试划分，以及 VenusX 留出的 44 个 InterPro 家族中的所有经审核成员进行筛查。如果某条序列与其中一条序列完全匹配，或通过 MMseqs2 命中其中一条序列，且两条序列的序列一致性至少为 30%、任一序列的覆盖度至少为 80%，则将其删除（easy-search，灵敏度 7.5，E 值为 10^{-3}）。这一步移除了 365,157 条序列，保留 273,160 条，其中 205,016 条在 Swiss-Prot 中有条目可供提取标签。其余序列在相同阈值下按 MMseqs2 簇进行划分，并删除具有训练命中的验证和测试蛋白质（7,141 个），最终得到 190,115 个训练蛋白质、2,413 个验证蛋白质和 5,347 个测试蛋白质。在训练蛋白质中，178,457 个带有功能、基因本体或 EC 注释，61,838 个至少带有一个残基级标签（其中 26,450 个带有活性位点，50,319 个带有结合位点）。GO-GPT 测试集未参与此次筛查，因为该筛查早于该评估；GO-GPT 的训练划分是五个来源之一，因此在该基准测试中仅排除了完全匹配的测试序列。

任务微调。 每个第 4 阶段模型都使用其基准测试自身的训练集进行训练，且不再进行序列一致性筛查。对于 GO-GPT 和 Mol-Instructions，通过精确匹配移除蛋白质出现在测试集中的训练数据。对于 Prot2Text，我们使用其发布的划分，不做任何改动（248,315 个训练蛋白质、4,172 个验证蛋白质和 4,203 个测试蛋白质；其作者使用 CD-HIT 按 40% 序列一致性构建），其中训练集与测试集之间没有共享序列；长度超过 1,024 个残基的序列会被截断。Prot2Text 的微调从 95k 更新检查点开始，并从 166k 更新起进行加权随机有放回抽取，使重复描述被采样的频率降低。由于这些划分均无法保证与训练集之间存在距离，我们按测试蛋白质与训练集的序列一致性分层报告每项功能结果。

A.4 蛋白质训练后的语言能力

表 2：**相对于基础版 Qwen3-8B 的准确率变化，以百分点计**。 使用完整测试集进行直接作答，另有说明的情况除外。推理模式 仅在包含 200 个项目的子集上启用（区间约为 ± 6 个百分点）。仅回放 指仅使用回放目标训练约 44k 步、未使用蛋白质数据的基础模型；其采用少样本提示进行评分（ARC-C 为 25 个，MMLU 和 GSM8K 为 5 个，TruthfulQA 为 0 个），因此只能在数值尺度上与其他数据行进行比较。其 TruthfulQA 得分在不同检查点之间会有数个百分点的波动。

模型	ARC-C	TruthfulQA	MMLU	GSM8K
RumiFold（95k 次更新）	−17.0	−19.4	−14.2	−21.1
基因本体微调后	−23.4	−31.6	−18.6	−27.3
经过 Prot2Text 微调后	−11.9	−25.0	−10.4	−13.2
RumiFold（95k 次更新），推理模式	−14.5	−21.0	−23.8	−14.0
仅回放，不包含蛋白质	−4.7	−3.8	−7.4	−8.7

任务微调会使成本朝两个方向变化：基因本体微调会在每个基准测试上增加成本，而 Prot2Text 微调则会降低除 TruthfulQA 之外所有基准测试上的成本（表 2）。

B 结构问题

问题类型与受扰动的折叠。第 2 阶段的结构问题涉及两个残基之间的距离、一个残基被埋藏的程度、距离另一个残基 8 埃以内的残基、形成远程接触或界面的残基，以及模型对某个片段折叠的置信度。每个标签都根据模型自身对该蛋白质预测的坐标计算得出。

从第 2 阶段后期开始，四分之一的结构问题会针对一个经过刻意降质的折叠过程提出：在折叠过程中向 ESM-C 输入加入高斯噪声，并调整噪声规模，使受扰动后的 lDDT 保持在 0.4 到 0.7 之间，然后根据生成的坐标计算标签。引入这一做法，是因为早期分析显示，部分答案是直接从编码器特征中读取的；它迫使答案与实际生成的折叠过程保持一致，而不是与该序列通常会产生的折叠过程一致。

B.1 使用与不使用折叠读取器

表 3：**使用折叠模型和不使用折叠模型作为蛋白质读取器进行结构问答**。使用较早的折叠检查点作为读取器时，在留出蛋白质上的精确答案准确率。两种设置均使用 Qwen3-8B、ESM-C特征和相同的问题集；不使用折叠的设置通过全新的桥接模块读取 ESM-C。需要三维几何的问题能区分两种设置；可依据局部序列上下文回答的问题则不能。

问题类型	不使用折叠	折叠
依赖几何信息		
两个残基之间的距离（值）	0.290	**0.704**
与给定残基距离最近的残基	0.045	**0.444**
某个残基的接触伙伴	0.079	**0.527**
四个候选项中最近的一个	0.361	**0.828**
远程接触（是/否）	0.722	**0.888**
链配对	0.345	**0.740**
可由序列读取		
溶剂暴露	0.833	0.844
二级结构状态	0.805	0.816
埋藏的螺旋残基	**0.661**	0.550
标题（各类型的均值）	0.484	**0.642**

表 3 比较了两个在相同结构问题上训练相同步数的问答模型：一个模型的蛋白质读取器是一个折叠检查点，另一个模型则通过一个全新初始化、未经折叠预训练的桥接模块读取 ESM-C 嵌入。对于可以从局部序列上下文中回答的问题，例如溶剂暴露或二级结构，这两个模型难以区分。对于需要三维排列的问题，两者差距很大：在判断两个残基之间的距离这一问题上，准确率分别为 70% 和 29%；在命名距离最近的残基这一问题上，分别为 44% 和 5%。ESM-C 嵌入包含大量结构信息，但在这样的训练预算下，连接到这些嵌入上的语言模型无法学会提取成对几何结构；而经过折叠训练的语言模型则可以直接利用这些信息。

表 4：**当前模型上的结构问答，以及模型读取信息时所用的注意力头**。针对 24 条留出链上每个模型自身的预测结构提出 264 个问题（144 个 Cα 距离问题、72 个较近片段问题和 48 个残基身份问题）。仅在读取问题时将注意力头置零；折叠保持不变。共享文本匹配接触头是我们检查的全部八个检查点中均存在的接触头，同时也能匹配基础版 Qwen 中的重复词元。随机头从相同的层中抽取（五次抽样的均值）。

	距离		较近者	残基
读取时置零的注意力头	r	平均绝对误差（埃）	片段	序列一致性
RumiFold（*95k* 更新次数）				
无（完整）	0.93	2.6	0.92	0.96
7 个共享文本匹配接触头	0.79	6.6	0.81	0.96
7 个随机头	0.93	2.6	0.91	0.96
19 个文本匹配接触头	0.64	8.1	0.79	0.96
19 个随机头	0.93	2.7	0.91	0.96
3 个共享接触头，不匹配文本	0.93	2.6	0.92	0.96
第 1–8 层中的 17 个文本匹配头	0.94	2.6	0.93	0.96
Structural-QA 检查点（*33k* 次更新）				
无（完整）	0.95	2.4	0.93	0.90
7 个共享文本匹配接触头	0.79	5.9	0.86	0.94
7 个随机头	0.95	2.3	0.92	0.91
19 个文本匹配接触头	0.67	6.8	0.78	0.94
19 个随机头	0.94	2.3	0.93	0.93
3 个共享接触头，非文本匹配	0.95	2.2	0.93	0.90
第 1–8 层中的 17 个文本匹配头	0.94	2.6	0.93	0.90

当前模型上的结构问题。 表 4 在 RumiFold 本身以及较早的 Structural-QA 检查点上重复了评估，问题涉及各模型自身的折叠。RumiFold 对 Cα 距离的陈述均值误差为 2.6 埃（与其折叠中的距离相关性为 0.93），在两段片段中选择较近者的准确率为 92%，并且在 96% 的情况下说出某一位置上的残基。

相同的评估显示，几何答案会经过哪些注意力头。我们检查的每个检查点中都有 7 个注意力头属于接触头；在基础版 Qwen 读取文本时，它们也都是从某个词元关注到该词元较早副本的注意力头（D节）。仅在读取问题期间将这 7 个注意力头置零，而保持折叠不变，会使 RumiFold 的距离误差升至 6.6 埃，比将同一层中的 7 个随机头置零高出 4.1 埃（95% 区间 $[3.4, 4.8]$），并使较近片段的准确率下降 11 个百分点（$[4, 18]$）。将它们的输出替换为其均值而非置零，也得到相同结果（$+3.7$埃，$[3.0, 4.4]$）。残基序列一致性没有变化：模型仍然能够读取蛋白质，但失去了成对几何结构。两组对照设置——不进行文本匹配的共享接触头，以及早期层中从未成为接触头的文本匹配头——都使所有答案保持在完整模型的水平。

C 更多折叠结果

CAMEO22 和 CASP14。 这两个基准测试的目标，以及与其序列一致性在30%以内的所有序列，均已从我们的训练数据中移除。我们使用了每个 CAMEO22靶标 [Robin等人， 2021] (183) 以及每个具有公开坐标的 CASP14靶标 [Kryshtafovych 等人， 2021] (34)，且不设长度上限（图10和表5）。在 CAMEO22 上，RumiFold 达到 0.805 的 TM-score、0.800 的 CA-lDDT 和 0.753 的 GDT-TS，优于 ESM3-open 和 DPLM-2，但落后于 SimpleFold 700M（0.816、0.846、0.786）。在难度更高的 CASP14靶标上，其指标分别达到 0.522、0.580 和 0.451：优于 ESM3-open、DPLM-2 和 SimpleFold 100M（0.491、0.569、0.427），但落后于 SimpleFold 700M（0.575、0.632、0.520）。SimpleFold 3B 在 CASP14 上进一步领先（0.637、0.685、0.589），超过 ESMFold（表5）。在两项测试中，RumiFold-Struct 均略优于 RumiFold（TM-score 分别为 0.814 和 0.533）。

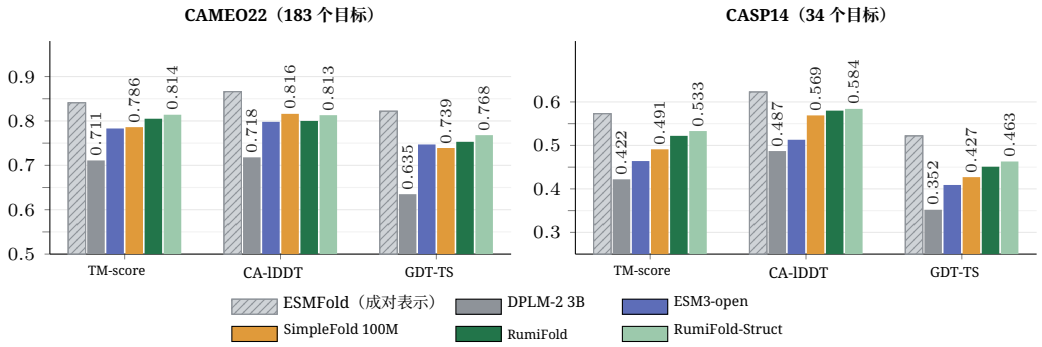


Figure 10：在训练中排除的两个基准测试上进行单序列结构预测 a t 30% 序列一致性，涵盖所有目标和所有长度，采用图3的布局。

在 CAMEO 2025（这三组中的第三组）上，SimpleFold 3B 在每项指标上都优于 RumiFold（TM-score 为 0.936，对比 0.893）。

与其他技能的折叠较量。RumiFold 最后经过了一个 30k 更新次数的阶段，训练权重偏向结构（88% 的结构项目）。它将 CAMEO22 的 lDDT 从 0.795 提高到 0.806，并将 CASP14 的得分从 0.569 提高到 0.574（在训练器的评测集上）；结构问答表现也有所提升（从 0.81 提高到 0.87），而功能指标则有所下降：GO 头的 F_{max} 从 0.450 降至 0.416，活性位点 AUPR 从 0.848 降至 0.736。这些技能共享一组适配器，并相互争用。

复合物。 RumiFold 接受两条链，因此可直接预测复合物；ESMFold 则可通过多聚甘氨酸连接肽实现这一点。本节结果针对 RumiFold-Struct。表 7 报告了 FoldBench 中预测结果达到可接受水平的界面比例（DockQ ≥ 0.23 [Basu 和 Wallner， 2016]），图 11 展示了一个示例。

表5：在 allCAMEO22（183 个）和 CASP14（34 个）目标上进行的单序列结构预测，不设长度上限：TM-score / Cα-lDDT / GDT-TS，完整链，使用我们的评分器。Chai-1 和 ESMFold 使用成对表示。

模型	CAMEO22	CASP14
ESMFold	.841 / .866 / .822	.573 / .623 / .522
Chai-1（单条序列）	.835 / .871 / .818	.566 / .629 / .523
SimpleFold 3B	.820 / .849 / .796	.637 / .685 / .589
SimpleFold 11亿参数	.818 / .843 / .786	.575 / .633 / .515
SimpleFold 700M	.816 / .846 / .786	.575 / .632 / .520
SimpleFold 100M	.786 / .816 / .739	.491 / .569 / .427
ESM3-open	.783 / .798 / .747	.464 / .513 / .409
DPLM-2 3B	.711 / .718 / .635	.422 / .487 / .352
DPLM-2 650M	.672 / .676 / .589	.385 / .460 / .320
RumiFold	.805 / .800 / .753	.522 / .580 / .451
RumiFold-Struct	.814 / .813 / .768	.533 / .584 / .463

表 6：使用基准测试作者的脚本评估的三个近期单链基准测试：所有目标的 Cα TM- 分数 / lDDT / GDT-TS。ESMFold 和 Chai-1（上方区块）使用成对表示。最后一条分隔线以上的所有模型均基于单条序列运行，不使用多序列比对或模板；AlphaFold3 和 Boltz-2（下方区块）则使用多序列比对，其结果仅作为基于多序列比对的预测所能达到水平的参考，并非同类比较。我们的训练数据仅针对 CASP14 和 CAMEO22 进行了筛查，因此无法保证这些数据集中的任何一个都未被用于 RumiFold 的训练。

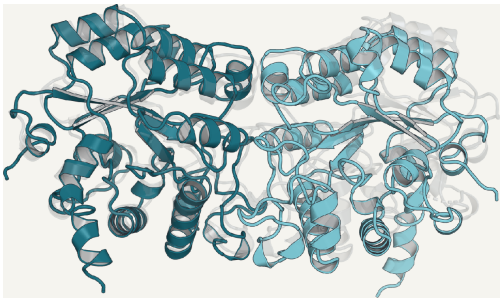
模型	FoldBench（334）	CASP16（155 个结构域）	CAMEO 2025（200）
ESMFold	.871 / .850 / .827	.764 / .767 / .710	.905 / .891 / .863
Chai-1	.871 / .853 / .827	.758 / .770 / .707	.921 / .910 / .885
SimpleFold 3B	.840 / .807 / .780	.706 / .727 / .647	.936 / .919 / .902
SimpleFold 700M	.832 / .802 / .765	.706 / .730 / .641	.928 / .914 / .889
SimpleFold 100M	.796 / .765 / .711	.662 / .702 / .589	.897 / .877 / .835
ESM3-open	.783 / .747 / .710	.651 / .668 / .588	.849 / .819 / .785
DPLM-2 3B	.709 / .659 / .591	.547 / .592 / .466	.802 / .744 / .684
DPLM-2 650M	.656 / .614 / .536	.515 / .562 / .436	.757 / .703 / .635
RumiFold	.840 / .781 / .756	.714 / .706 / .636	.893 / .833 / .818
RumiFold-Struct	.846 / .792 / .767	.720 / .714 / .642	.896 / .848 / .830
使用多序列比对			
AlphaFold3	.937 / .927 / .914	.858 / .865 / .818	.960 / .948 / .932
Boltz-2	.935 / .927 / .909	.846 / .857 / .801	.961 / .953 / .938

C.1 语言模型为折叠带来了什么?

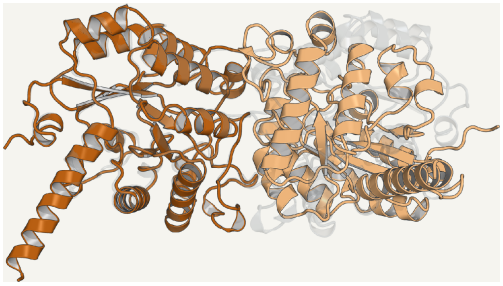
RumiFold 从一个经过文本预训练的模型开始。这样的预训练对折叠有帮助吗？还是说，任何同等规模的 Transformer 都能胜任？我们在小规模条件下（约 2k 个训练蛋白质、两个随机种子）进行了受控比较，结果取决于训练设置。

在冻结基础模型的情况下，也就是 RumiFold 所采用的训练设置中，预训练权重确实有帮助。经过 4k 次更新后，使用低秩适配器的预训练 Qwen 比以相同数量的可训练参数从头训练的 Transformer 高出 0.054 lDDT $[0.042, 0.069]$；也比将所有冻结的权重矩阵随机打乱的 Qwen 高出 0.075 $[0.058, 0.093]$（图 12a）。两种随机种子下的效果大小有所不同，因此我们将其视为对这一趋势的支持，而非对效果大小的支持。

当所有权重都参与训练时，预训练带来的优势只是抢先起步。文本预训练能让模型在训练早期更快学习：在 4k 次更新时，lDDT 高出 8 到 9 分；但到了 16k 次更新时，这一优势便消失了，而且在两种随机种子下，发布的 Qwen 权重最终都不如随机初始化（图 12b）。预训练语言模型为折叠带来的是一种适配器可以利用的权重组织方式，而不是经训练后仍能保留下来的知识。



(a) RumiFold-Struct：DockQ 0.77，界面均方根偏差 1.7 埃



(b) ESMFold：DockQ 0.04，界面均方根偏差 12.3 埃

图 11：**一个留出同源二聚体（PDB 8JQF，** 2×305 **个残基）**。天然结构以灰色显示；每个预测结果都与其左侧链叠合，因此右侧链显示了互作伙伴被放置的位置。ESMFold 对每条链的折叠预测更准确（lDDT 0.96 对比 0.83），但对接位置有误。最接近的训练序列有 40% 的序列一致性。

表 7：FoldBench 界面上的对接成功率（DockQ ≥ 0.23）。留出条件：每条链与 RumiFold-Struct 训练链的序列一致性均低于 50%。

界面类型	n	RumiFold-Struct	ESMFold（连接子）
蛋白质–蛋白质	244	**43%**	19%
蛋白质–蛋白质，留出测试	50	**32%**	26%
抗体–抗原	153	1%	0%
蛋白质–肽	36	**19%**	14%

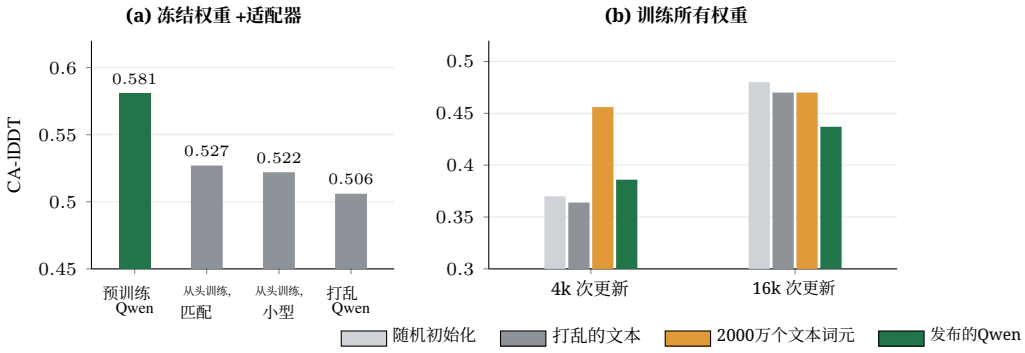


图12：**当基础模型被冻结时，文本预训练会带来帮助；而在模型未被冻结时，这种帮助只在训练早期出现。**小规模折叠实验，每种设置使用两个随机种子，使用开发集蛋白质。**(a)** 采用我们的方案（冻结基础模型，使用秩为64的适配器），经过4000次更新后：将预训练的Qwen与从头训练的主干网络，以及对每个权重矩阵进行打乱的Qwen进行比较。**(b)** 采用Qwen架构并训练所有权重，使用四种初始化方式。文本预训练带来的领先优势到16000次更新时便已消失。柱形表示两个随机种子的均值；各随机种子的值和区间见正文。

D 接触头承载折叠

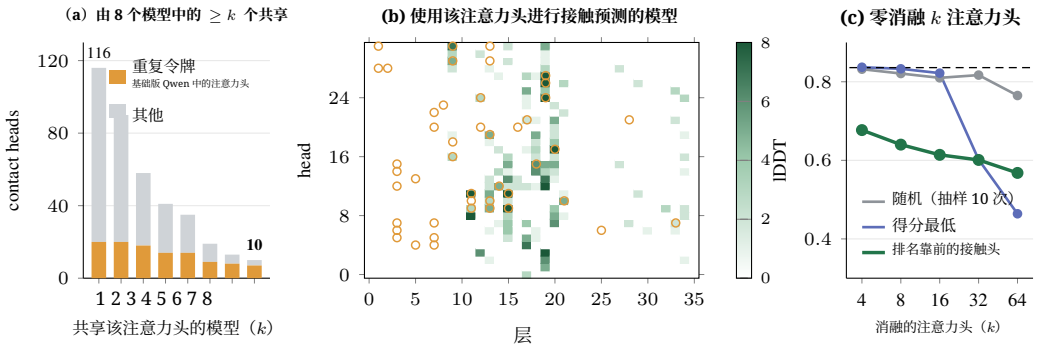


图 13：**分别训练的模型使用相同的注意力头进行接触预测，其中大多数注意力头在基础版 Qwen 中对应重复词元。（a）**在八个检查点（RumiFold 系列中的五个，分别训练的三个）中，至少在 k 个检查点里充当接触头的注意力头数量。十个注意力头在全部八个检查点中都是接触头，其中十个里的七个在基础版 Qwen 中是重复令牌头（琥珀色）。**（b）**对于这 36×32 个注意力头，显示其在多少个检查点中充当接触头；圆环标记出其中 45 个注意力头：在基础版 Qwen 读取文本时，它们会从某个词元关注到该词元较早出现的副本。**（c）**将 k 个注意力头的输出置零后，RumiFold-Struct 的 lDDT（虚线：未修改，0.836）。移除表现最好的四个接触头会使 lDDT 降低 0.16；移除四个随机头则毫无影响。

我们根据 RumiFold-Struct 在折叠过程中产生的注意力图与天然接触图（Cα，距离在 8埃以内）的匹配程度，对其每个注意力头进行评分。评估使用 24 条留出的 CATH 链（每条最多 160 个残基），计算所有间隔至少六个残基的残基对中，注意力权重与接触情况的曲线下面积，以及该注意力头排名前 L 的残基对中间隔至少 24 个残基的残基对的精确率。

接触头出现了。 折叠训练后，有 39 个注意力头的接触曲线下面积高于 0.85，主要集中在第 9 至第 20 层，此外第 33 层和第 34 层还有第二组注意力头。关闭适配器后，同一网络在相同输入上的任何注意力头都不再表现出这一特征（最高曲线下面积为 0.67）。这与蛋白质语言模型的接触头 [Rao 等人， 2020], 相呼应，不同之处在于 Qwen 从未在蛋白质上进行预训练：这种行为完全由低秩更新注入。

它们是必需的。 将得分最高的四个接触头归零，会使 lDDT 从 0.836 降至 0.677（图 13c）。将四个随机注意力头归零后，平均值仍为 0.832，十次随机抽样中没有一次低于 0.805。RumiFold 也呈现相同结果（从 0.826 降至 0.444），其基因本体微调版本也是如此（从 0.851 降至 0.795）。当注意力头达到 32 个或更多时，排序发生反转：移除得分最低的注意力头，其影响不亚于移除得分最高的注意力头，甚至更严重，这说明许多注意力头会以接触得分无法捕捉的方式参与蛋白质折叠。

借鉴自文本。在预训练于文本的模型中，接触头从何而来？我们比较了八个检查点，其中五个来自 RumiFold 的发展脉络，另外三个则从全新初始化开始分别训练。每个模型都有 29 到 75 个接触头，而且这些头大多相同：八个检查点中都有十个接触头（图 13a、b）。这十个头中，有七个在基础版 Qwen 阅读文本时，会从一个词元注意到同一词元较早出现的副本；若是随机注意，概率则为 0.4。从不成为接触头的，是那些注意前一个词元的头。这七个头不可或缺。将它们归零，会使仅经折叠训练的模型（包括单独训练的模型）的 lDDT 降低 0.18；而在 RumiFold 及其微调版本中，降低幅度为 0.03 到 0.04。相比之下，将同一层中的七个随机头归零，造成的降幅至多为 0.002。它们还承担结构问答功能（表 4）。折叠微调似乎重新利用了基础模型用于寻找词元较早副本的注意力头：“找到相同的词元”变成了“找到与之接触的残基”。

E 更多功能基准测试

BioReason-Pro 基准测试。第 5 节中的基因本体结果使用 GO-GPT v2 基准测试。BioReason-Pro [Fallahpour 等人、 2026] 以及首个 GO-GPT 均是在较早的基准测试上训练和评估的

。为了按照各自的评估标准进行比较，我们在该基准测试自己的训练集上训练 RumiFold，并使用该基准测试自己的评分器，在其测试集上对模型进行评分。其他设置均保持不变：训练从与主模型相同的 95k 次更新检查点开始，使用相同的基因本体训练方案，在 120,143 个训练蛋白质上训练 13 轮。测试集包含 BioReason-Pro 发布版本中的 8,159 个蛋白质，评分采用该发布版本评估脚本计算的信息加权 F-max，并对三个方面的结果取平均。

表 8：**在 BioReason-Pro 自有数据上训练的 RumiFold，在 BioReason-Pro 和 GO-GPT 的测试集上表现更佳**。在 8,159 种测试蛋白质上的信息加权 F-max，按方面列出并计算平均值，每行使用相同的评分器。粗体表示各列中的最佳模型。

模型	MF	BP	CC	Mean
BioReason-Pro（监督微调）	0.712	0.515	0.685	0.638
BioReason-Pro（强化学习）	0.716	0.543	**0.726**	0.662
GO-GPT v1	**0.717**	0.550	0.725	0.664
RumiFold ，在相同数据上训练	0.704	**0.589**	**0.726**	**0.673**

RumiFold 得分为 0.673，而 GO-GPT 为 0.664，采用强化学习的 BioReason-Pro 为 0.662，其监督版本为 0.638 （表 8 ）。优势体现在生物过程中（ 0.589，而 GO-GPT 为 0.550 ）；在细胞组分上持平，在分子功能上则落后 0.013。

Mol-Instructions。 Mol-Instructions [Fang 等人 2023] 要求根据四项指令（蛋白质功能、一般功能、催化活性、结构域和基序），用一句话回答有关某种蛋白质的问题，并通过 ROUGE-L [Lin, 2004]与基于 Swiss-Prot 得出的参考答案进行比较评分。我们在其训练项目上对 RumiFold 进行微调，并使用贪心解码，在包含 9,128 个项目的完整测试集上进行评估。匹配对照组 *Qwen + ESM-C* 使用相同的语言模型、相同的 ESM-C特征、数据和步数，但未进行折叠预训练。

表 9：**Mol-Instructions**。 在所有测试项目，以及与训练集序列一致性低于 50% 和 30% 的项目上的 ROUGE-L $\times 100$ 。RumiFold、其对照模型和基准下限均在完整测试集上进行评分（9,128 个项目；n=2，两个子集分别包含 433 个和 816 个项目）。已发表系统(†)在其中的一个样本上运行：即所有系统都给出答案的 1,790 个项目（n=465 和 146 个）；在该样本上，RumiFold 的得分为 65.6 / 55.5 / 45.3。下限 行并非训练模型。粗体表示每列中的最佳得分，下划线表示第二佳得分。

系统	All	<50%	<30%
RumiFold	<u>66.2</u>	<u>56.4</u>	**45.9**
Qwen + ESM-C，无折叠	63.1	52.0	<u>42.7</u>
Mol-Instructions LLaMA-7B† [Fang 等人， 2023]	53.0	42.1	36.4
EvoLlama† [Liu 等人， 2024]	48.6	38.9	33.4
Prot2Chat† [Wang 等，2025b]	35.6	32.6	27.2
TourSynbio† [Shen 等人， 2024]	17.7	18.4	20.3
InstructProtein† [Wang 等人，2023b]	16.1	15.5	15.4
Evolla-10B† [Zhou 等人， 2025]	13.0	12.2	10.8
BioMedGPT-10B† [Luo 等人， 2023]	9.9	9.9	7.9
最近的训练蛋白质（下限）	**67.3**	**56.7**	42.5
模板（下限）	40.7	40.4	38.8

RumiFold 得分为 66.2，而其匹配对照组为 63.1（表 9）：折叠预训练总体上带来 3.1 分的提升 $[2.8, 3.5]$，在序列一致性低于 50% 时提升 4.4 分 $[3.7, 5.2]$，低于 30% 时提升 3.2 分 $[1.7, 4.7]$。这一优势在四条指令下均成立。在经过训练的蛋白质模型中，RumiFold 在该基准测试上达到了最先进水平：已发表系统中表现最佳者，在我们运行该系统的样本上得分为 53.0，而 RumiFold 得分为 65.6。直接复制最近的训练蛋白质的答案仍然是一个很强的参照，因为大多数测试蛋白质都有一个几乎完全相同的训练邻居：其总体得分为 67.3，在序列一致性低于 50% 时与 RumiFold 持平，而在低于 30% 时落后于 RumiFold（42.5 对 45.9）。

表10：**功能清晰可辨之处**。 在每个表征空间中进行最近邻基因本体查找（k=20），从 15,000 个训练蛋白质到测试集；加权 F_{max}。ESM-C 取自其最后一层；每个模型取自其最佳层。纯度：蛋白质的 10 个最近邻中，属于同一顶层分子功能类别的比例。

表征	All	<30% 序列一致性	纯度
ESM-C，最后一层（模型的输入）	0.435	0.363	0.70
Qwen + ESM-C，经过基因本体微调（未进行折叠预训练）	0.427	0.351	0.73
RumiFold-Struct（仅折叠）	0.455	0.372	0.79
RumiFold-Struct ，基因本体微调	**0.523**	**0.420**	**0.90**

F 折叠后的表示

第5.1节的对照表明，折叠预训练会影响得分。这里我们考察表征本身。如果学习折叠能教会语言模型描述蛋白质，那么，与它获得的 ESM-C特征相比，从其状态中读出蛋白质属性应当会有更好的效果，而功能训练应能进一步改善这些表征。我们使用冻结状态和简单的读出进行测试：最近邻和小型探针。在全文中，ESM-C 都取其最后一层的输出，这正是提供给模型的内容。

可以通过状态查找功能。表10探讨了在各个表征空间中，通过最近邻查找功能的效果如何。仅经过折叠训练的模型，其状态比它读取的特征更有用（总体得分为 0.455，对比 0.435；邻居共享同一功能类别的比例为 79%，对比 70%），尽管它从未见过任何功能标签。随后，功能微调进一步提升了折叠预训练模型的表现（0.523）；而对普通 Qwen 进行相同的微调后，其状态的实用性仍低于它原本的特征（0.427）。在更大的测试集上再次进行查找也得出了相同结论：对于分子功能，仅折叠状态中的邻居得分为 0.462 F_{max}，而 ESM-C特征中的邻居得分为 0.391。

残基层面探针。第 6.1 节总结了在 VenusX 上进行的残基层面探针结果；表 11 列出了所有任务和数据划分。表中还列出了 ESM-2 的最后一层作为参考。在这些残基任务上，ESM-2 是更强的编码器：它的两种规模在每项活性位点、结合位点、保守位点和基序任务上的表现都优于 ESM-C 和仅折叠状态；只有在三个结构域任务上，ESM-2 的表现才落后于仅折叠状态。

表中还给出了未进行折叠预训练的对照组：在基因本体上微调的*Qwen + ESM-C* 。在全部 17 项任务中，它的状态都不如输入给它的 ESM-C 特征适合读取残基：在 MF50 和 MP50 划分中的活性位点、结合位点、保守位点和基序任务上，AUPR 低 0.11 至 0.30；其他任务上则低 0.01 至 0.09。在 17 项任务中，有 15 项低于仅折叠状态；在跨家族保守位点任务上持平，在跨家族活性位点任务上则领先（0.144 对 0.101）。将 ESM-C 输入仅经过功能训练的语言模型，会丢失折叠目标所保留的残基层级信息。

功能训练会进一步优化这些表示。自由文本功能也有同样的结果。在 Prot2Text 测试集上，复制最近的训练蛋白质在 ESM-C 最后一层中的描述，BLEU-4 得分为 33.5；在未使用测试集的情况下选定深度后，在 Prot2Text 训练后的 RumiFold 状态中进行同样的查找，得分为 44.6。同一模型第一层的输入得分为 37.7，因此大部分提升是在语言模型内部实现的。

F.1 隐状态中的结构

依赖关系中的接触。 我们直接考察这些状态是否具有类似蛋白质的结构。对于40种测试蛋白质，我们依次遮蔽每个残基，并记录其他每个残基的状态变化幅度。由此得到的依赖关系图还原了蛋白质的接触图：在第23层之后最强的远程依赖关系中，54%是真实接触，而基线比例仅为1.6%（图14a）。功能微调使这一特征一直保持到第23层，只改变后续层；在这些层中，状态的响应性更强，也不再那么受接触模式支配。网络中部保留了结构；网络末端则针对任务进行了重塑。

表11：VenusX 上的残基层面探针：每个模型的最后一层。对每项任务和数据划分（MF50、MP50 和跨家族），在各编码器的冻结特征、RumiFold-Struct 的第二轮最终状态以及*Qwen + ESM-C*（不进行折叠、在基因本体上微调的对照模型；第5.1节）的最后一层状态上训练同一探针，并报告汇总测试集 AUPR。粗体表示 ESM-C 与 RumiFold-Struct 中较高者；ESM-2 列仅供参考。

任务、数据划分	ESM-2 650M	ESM-2 3B	ESM-C-6B	Qwen +ESM-C	RumiFold-Struct
活性位点 , MF50	.873	.860	.707	.488	**.789**
活性位点 , MP50	.944	.944	.864	.679	**.922**
活性位点，跨家族	.154	.154	**.151**	.144	.101
结合位点（InterPro），MF50	.933	.923	**.842**	.643	.804
结合位点（InterPro），MP50	.966	.966	.918	.761	**.923**
结合位点（InterPro），跨家族	.292	.327	.252	.167	**.269**
结合位点（BioLiP），MP50	.399	.405	**.395**	.315	.390
保守位点 , MF50	.902	.912	.788	.491	**.831**
保守位点 , MP50	.908	.922	.792	.505	**.818**
保守位点，跨家族	.247	.261	**.261**	.209	.209
基序 , MF50	.899	.917	**.812**	.620	.786
基序 , MP50	.849	.849	.780	.666	**.802**
基序，跨家族	.486	.514	.433	.364	**.436**
领域 , MF50	.613	.615	.578	.502	**.640**
领域 , MP50	.618	.630	.598	.533	**.654**
领域，跨家族	.513	.471	.593	.517	**.637**
表位，MP50	.167	.170	**.179**	.149	.177

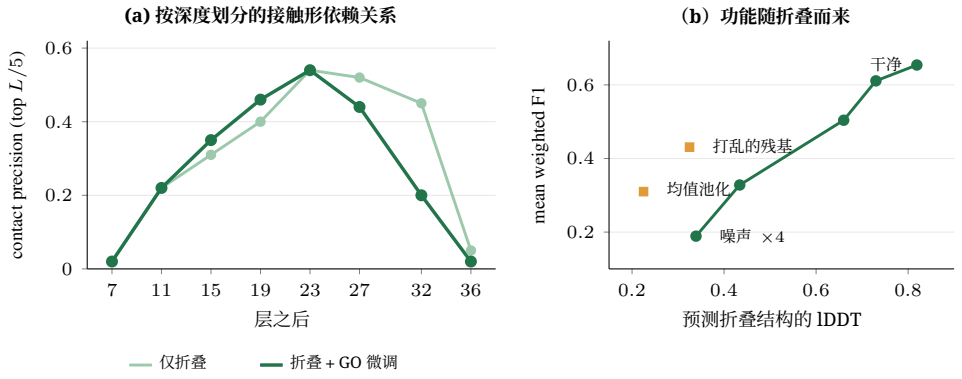


图 14：**语言模型内部的结构及其与功能的联系。（a）**依次遮蔽输入中的每个残基，并记录其他每个残基状态的变化；将所得图与真实接触图进行比较（40 种蛋白质；随机情况下为 0.016）。接触形依赖关系从第 11 层开始出现，并在第 23 层达到峰值，而功能微调并未影响这一峰值。**（b）**降低语言模型输入的质量，会同时降低折叠准确率和功能预测准确率（RumiFold-Struct + 基因本体 GO 头；曲线：仅进行一次前向传播，然后逐渐增加噪声）。

功能取决于折叠。 破坏折叠会损害功能预测。随着语言模型的输入逐渐退化，lDDT 和加权 F1 同步下降，从0.82和0.65降至0.34和0.19（图14b）。打乱残基会破坏折叠（降至0.33），但保留了更多功能信息（0.43）；这符合预期：部分功能信号来自组成，其余则来自排列方式。

F.2 富集是在折叠过程中计算的

我们限制每个残基在折叠轮次中沿链能够看到的范围，同时保持编码器的特征不变，并询问模型其蛋白质具有的某项功能。当没有任何残基能够看到另一个残基时，模型能识别出 18% 的蛋白质的功能；随着可见范围扩大，答案逐渐恢复：达到八个位置时为 62%，达到三十二个位置时为 85%，覆盖整条链时则为 100%。这种能力也会随着轮次的深度逐步建立：将较早层的状态传递给文本时，分子功能 F1 在输入处为 0.39，在第 24 层为 0.55，而完整模型为 0.67。此外，折叠还会将活性

位点的残基连接起来：扰动一个催化残基处的输入，会改变其互作伙伴中存储的信息；该伙伴通常距离其 145 个位置，这种改变幅度是空间中距离相同的其他残基的两倍（0.17 对 0.08；G节）。功能并不是从编码器中直接传递过来的，而是模型在折叠过程中推导出来的。

G 折叠工作空间

本节给出图 6b 背后的完整结果，并逐一介绍各个酶。该模型是经过 GO 训练后的 RumiFold（第 187,550 步）。催化残基是 UniProt 的活性位点注释，并向两侧各延伸两个残基，因此其中没有任何内容是针对该模型拟合的。问题是“该蛋白质是否具有 GO 分子功能术语‘催化活性’？”，这些蛋白质是留出酶；在未编辑时，它们的答案为“是”。

表12：**突变、恢复、阻断：所有条件**。在图 6b 中 1,448 个酶里长度为 80 至 500 个残基且运行了每种条件的 1,037 个酶中，声称自身具有催化活性的留出酶所占比例（括号内为 95% 自助法区间）。最后一列仅限于 339 个答案因突变而改变的酶。读取器是最关注该蛋白质催化残基的注意力头，在没有标签的情况下选出；被阻断的注意力头会接收其从突变蛋白质中读取的内容。

条件	所有酶（%）	因突变而发生变化（%）
未编辑	100	100
1. 催化残基突变为丙氨酸	67 [64, 70]	0
相同数量的随机残基发生突变	97 [96, 98]	93
2. 催化残基处发生突变，记忆已恢复	96 [95, 97]	90
其他残基处的记忆已恢复	67 [65, 70]	3
值已恢复，键仍处于突变状态	88 [86, 90]	64
键已恢复，值仍处于突变状态	73 [70, 76]	26
3. 已恢复 ，随后 32 个读取器被阻断	86	57
64 个读取器被屏蔽	79	36
128 个读取器被屏蔽	74	22
256 个读取器被屏蔽	70	11
256 个随机注意力头被屏蔽	93	80

所有条件。 表12列出了所有条件。在每一层，残基的记忆由两部分组成：键决定哪些注意力头会关注它，值则是这些注意力头读取的内容。恢复这两部分，可以挽回突变所改变答案的 90%；单独恢复值可以挽回 64%，单独恢复键则可以挽回 26%。屏蔽效果随数量变化：在 1,152 个注意力头中，屏蔽 32 个会使恢复的答案减少约三分之一，屏蔽 256 个则会使其减少近九成；而屏蔽 256 个随机注意力头只会使其减少约十分之一。

实例分析。 图15展示了枯草杆菌蛋白酶中的答案从何处读取；表13则追踪了四种酶在各个步骤中的情况。对每种酶而言，突变催化残基会使书面答案中不再出现该酶自身的活性，而恢复这些残基处的记忆则会使其重新出现。在抽取的 264 种酶中，有 48 种在每个步骤和对照中都呈现完整的肯定与否定答案模式，其中 13 种还以这种方式改变了书面答案。

局限性。 大多数酶在催化残基发生突变后，答案仍保持不变；与是或否的答案相比，书面列出的功能变化更少。单个替换的影响较小：UniProt 记录的 699 个会消除或降低活性的替换中，每次单独应用时有 12% 会改变模型的答案；而在 475 个记录为无影响的替换中，这一比例为 7%。模型对这些突变的敏感方向是正确的，但无法预测点突变体的表型。

折叠会连接活性位点的残基。 当输入在一个催化残基处受到扰动时，其最远催化伙伴处的记忆变化为 0.171 $[0.163, 0.180]$（相对变化），而空间距离匹配的残基处为 0.083。该效应在折叠过程的中间层达到峰值；当注意力仅限于 ± 8 个残基时，效应则不存在。

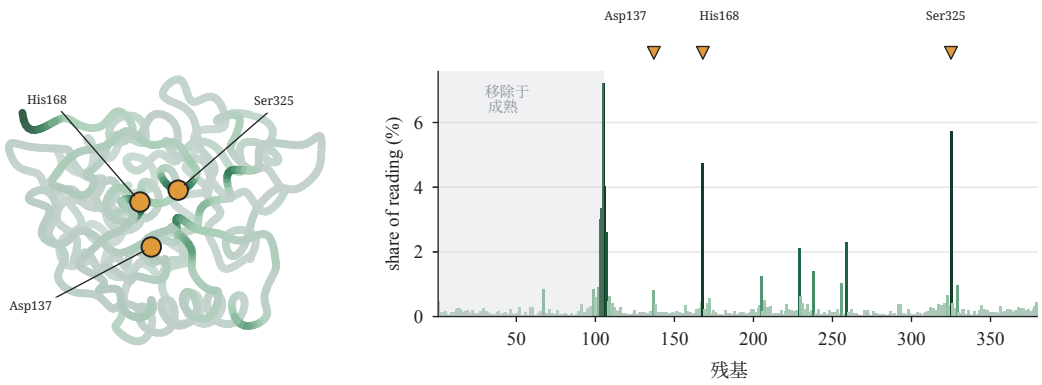


图15：**卡尔斯伯格枯草杆菌蛋白酶（P00780）**。*左*：成熟酶，按答案读取每个残基的程度着色，并标出了催化三联体。*右*：答案在379个残基的前体中每个残基上的注意力占比。His168和Ser325分别是读取量第三高和第二高的残基；读取量最高的片段是残基103至107，即前肽从成熟酶上切除的位置。该结构为 AlphaFold 数据库模型，仅用于展示。

表13：**四种酶，逐步进行**。模型对“这种蛋白质具有哪些分子功能？”的书面答案，分别针对未编辑的蛋白质，以及将催化残基及其邻近残基突变为丙氨酸后的蛋白质。在全部四种情况下，恢复这些残基处的记忆都会还原未编辑的答案；突变随机残基不会改变答案；阻断256个读取器则会使关于催化活性的答案变回“否”。这些被选为干净的案例。

酶（催化残基）	未编辑	催化残基突变
卡尔斯伯格枯草杆菌蛋白酶（Asp137、His168、Ser325）	蛋白质结合；丝氨酸型内肽酶活性	蛋白质结合
tRNA腺苷脱氨酶ADAT2 (Glu73)	tRNA特异性腺苷-34脱氨酶活性；蛋白质结合	蛋白质结合
双特异性磷酸酶21（Cys105）	蛋白酪氨酸磷酸酶活性；蛋白质结合	蛋白质结合
醛糖1-差向异构酶 (His205, Glu345)	醛糖1-差向异构酶活性	醛糖结合；β-N-乙酰氨基葡萄糖苷酶活性

G.1 通过值记忆追踪功能

本节给出表14背后的定义和进一步结果。所有结果均使用经过 GO 训练（第 187,550 步）后的 RumiFold，其文本过程没有编码器路径：每层的键/值行是它观察蛋白质的唯一视图。每次编辑都会改变值行或注意力头检索，而保持键不变。该协议在验证蛋白质上固定后运行一次，并在测试蛋白质上执行；答案采用贪心方式生成。

定义。对于一个术语 T，其方向是在每一层上，具有 T 的训练蛋白质值行均值减去不具有该项的训练蛋白质值行均值，并基于 16,000 个训练蛋白质进行拟合。一个注意力头的检索是其值行的注意力加权和；一个术语排名前 32 的注意力头，是指在 1,152 个注意力头中，其检索在 6,000 个训练蛋白质上最能区分 T 的 32 个注意力头。清零操作会将每一层中每个残基的值行都设为训练均值。这九个术语是分子功能术语，例如 ATP结合、钙离子结合、DNA 结合和蛋白激酶活性；锌（6 个蛋白质）、血红素（5 个）和甲基转移酶活性（4 个）的测试集规模较小。

一个坐标，以及它所经过的注意力头。表14给出了完整链路。清除记忆后，仅恢复第 8 层至第 24 层中沿 T 方向的每一行坐标，就能部分恢复功能（19% 的答案为“是”，T 在 30% 的答案中被写出）；而随机坐标不起作用。

表14：**从值记忆到答案**。 具有 GO 术语 T 的留出测试蛋白质，并且确实如此（215 种蛋白质，九个术语）。回答“是”：询问模型该蛋白质是否具有 T 时，模型回答“是”。写出 T：T 出现在其自由回答中（83 个答案）。“清除”是指用训练均值替换每一行值。括号内为 95% 自助法区间。

值记忆	文本头	回答“是”（%）	写出 T （%）
完整	完好	100	81
已擦除	完好	0	0
已擦除，T 坐标已恢复	完好	19 [13, 24]	30 [20, 40]
已擦除，随机坐标已恢复	完整	0	0
已清除，恢复了 T 坐标	T的前32个注意力头保持在其均值	6 [3, 9]	10
已清除，恢复了 T 坐标	保持32个随机注意力头	19	27
已清除	根据其干净检索结果，T的排名前 32 个注意力头	47 [40, 53]	57 [46, 66]
已擦除	根据其结果的 32 个随机头	2	2
保持完整	将 T 的排名前 32 个注意力头保持在其均值	86	77
已移除 T 坐标	完整	88	81

将 T 的排名前32的注意力头固定为其均值，会使这种恢复效果降至 6%；固定 32 个随机头后，恢复效果仍为 19%。对排名前32的注意力头进行检索会直接改变是/否对数几率（$+0.62$；相比之下，随机头为 $+0.08$ ）。

注意力头。 在彻底清除检索信息后，215 个蛋白质中仍有 13% 给出肯定回答。仅保留某个术语排名前 8 的注意力头时，这一比例达到 44%（随机 8 个注意力头为 13%）；保留排名前 32 的注意力头时达到 65%（随机情况下为 19%）。在这 32 个注意力头中，仅保留沿 T 方向的坐标时，比例为 63%，而随机方向仅为 12%。跨术语反复出现的注意力头，是通过对更早检查点进行解码分析独立发现的。

移植。 在每个层内按注意力排名匹配，将供体的值行复制到受体中，会使书面答案向供体的答案靠近（200 对测试样本中，两者没有共同的分子功能术语）：与供体答案相比，未编辑时的 F1 值为 0.25；每层使用 8 行时为 0.47，使用 32 行时为 0.52，使用全部数据行时为 0.56；同时复制键时则为 0.68；随机 32 行为 0.27。读取过程分布在不同深度上：仅使用第 8 至第 32 层就能重现完整效果，而不存在任何连续的 12 层区间可以做到这一点。同样的移植也会改变 Prot2Text 模型中的自由文本描述（与供体描述相比的 ROUGE-L 从 0.10 提升至 0.30，随机数据行为 0.10）。

局限性。 这些是对充分性的测试。移除的效果较弱：去掉 T 的坐标，或固定其 32 个注意力头，仍分别有 88% 和 86% 的蛋白质回答“是”，因此，一个类别方向只是众多载体之一。我们尝试的编辑都无法做到精确控制：只有在大幅改变答案其他内容的剂量下，每个方向才会写入或移除一个术语。模型推理中的残基级陈述（“在 Cys 14 处结合锌”）并非从其引用的数据行中读取：抹去这些数据行后，该陈述仍有 93% 的概率保留。

G.2 输入端的残基干预

如果模型的预测源自对蛋白质结构的表征，那么预测应取决于执行该功能的残基，以及它们在空间中的邻近残基。

功能位点。 我们选取带有已注释功能位点及其匹配 GO 术语的测试蛋白质，并在输入给语言模型的内容中遮蔽该位点的残基，同时保留注意力头对编码器的直接访问。在折叠预训练模型中，移除活性位点（中位数为 1 个残基）会使催化活性的 logit 降低 0.95，而移除相同数量的随机残基则毫无影响（表15）。金属配体和跨膜片段也表现出相同的情况。在对照组中，同样的干预没有影响，因为其预测并不经过语言模型。这种影响也体现在书面答案中：遮蔽活性位点时，RumiFold 会在 39% 的蛋白质上改变答案，而随机残基对应的比例仅为 4%，且这些变化具有特异性。DNA 甲基转移酶保留了“DNA 结合”，但失去了甲基转移酶相关术语；金属羧肽酶保留了“锌离子结合”，但失去了肽酶相关术语。

表15：**功能位点通过折叠路径发挥作用**。对于具有注释位点且匹配基因本体术语的测试蛋白质，仅将该位点的残基在语言模型的输入中屏蔽（注意力头直接查看编码器的部分不受影响）。上： 匹配术语的 logit 变化，括号中为相同数量随机残基对应的变化。下： 对于 RumiFold，书面答案发生变化的频率。

匹配基因本体术语的 *logit*	RumiFold-Struct + GO 头	Qwen + ESM-C
活性位点 →催化活性 (n=145)	-0.95 ($+0.01$)	-0.001
金属配体 →金属离子结合（n=35）	-0.99 （-0.02）	-0.05
跨膜片段 → 膜（n=210）	-4.08 （-0.76）	-0.22
RumiFold 的书面答案	答案已更改	不再写入功能
活性位点，中位数为一个残基（n=135）	39%	13%
数量相同的随机残基	4%	1%
与活性位点形成三维接触的残基（n=89）	16%	1%
与其相距较远的匹配残基	3%	0%
DNA结合区域（n=11）	91%	64%
随机片段	32%	5%

空间邻近残基。 更直接的结构检验，是考察在三维空间中靠近活性位点、但在序列上与之相距较远的残基。遮蔽约 7 个这样的残基，会使 16% 的蛋白质书面答案发生变化；而对于在埋藏程度和序列距离上匹配、但在空间上远离活性位点的残基，这一比例为 3%（$p = 0.001$）。

G.3 折叠记忆上的稀疏自编码器

6.2节中的干预会沿着我们选择的方向编辑记忆。稀疏自编码器则探究记忆自身具备哪些方向。我们在 RumiFold 进行基因本体训练前后的语言处理过程所读取的值（五层，第 8 层至第 24 层）上训练 TopK 自编码器 [Gao 等人， 2024] 其包含 16,384 个特征，每次激活 64 个；同时在原始 ESM-C 输入上进行训练以作比较。每个自编码器均使用训练语料库中的 1.04 亿个残基进行训练，并排除了基因本体验证蛋白质和测试蛋白质。

这些特征承载着答案所使用的信息。基因本体模型的自编码器在留出蛋白质上有 9% 的方差无法解释。将其重建值写回记忆，以替代原始值，对答案的影响很小：在 300 个验证蛋白质上，生成术语的 F1 值为 0.774，而未编辑时为 0.781。

这些特征代表什么。 我们根据 100,000 种蛋白质的 Swiss-Prot 注释为特征标注：562 个残基级概念（特征类型、配体、修饰、二级结构）和 8,714 个蛋白质级概念（家族、结构域、基因本体术语、酶类）。只有当某个特征在留出蛋白质上预测某个概念的表现优于同等规模的未训练字典时，才会为其赋予标签。按此标准，基因本体模型的特征中有 13.7% 获得标注，而未训练的字典中这一比例为 1.1%（表 16）。大多数获得标注的特征都能检测出某个蛋白质家族或结构域，并且只在该家族或结构域成员的少数残基上触发。约每一百个特征中，只有一个专门对应某类残基：结合泛醌的组氨酸（F1 值 0.96）、二硫键中的半胱氨酸（0.75）、氨酰-tRNA 合成酶的 KMSKS 基序（0.71）、信号肽（0.68）， 以及铁硫簇中的半胱氨酸（0.64）。

表 16：**已标注特征与概念覆盖度**。 已标注：通过标注规则的活跃特征所占比例。覆盖度：拥有专属特征的已注释概念数量；在 179 个领域、重复序列和基序概念、141 个结合位点概念以及 69 个活性位点概念中统计。

训练自编码器所用的数据	已标注（%）	结构域 / 基序	结合位点	活性位点
ESM-C 输入	21.3	37	17	7
GO 训练前的 RumiFold 记忆	13.6	50	32	13
GO 训练后的 RumiFold 记忆	13.7	**64**	**35**	**19**
未训练的字典	1.1	–	–	–

折叠带来了位点级特征。 ESM-C 输入的标注特征总数更多，其中大多数是家族检测器。折叠记忆中专门用于功能位点的特征更多：经过基因本体训练后，69 个活性位点概念中有 19 个各自对应一个特征，训练前则有 13 个；相比之下，ESM-C 中有 7 个。结合位点、结构域和基序也呈现相同的顺序（表 16）。其中大部分特征在进行任何基因本体训练之前就已存在。

局限性。 这些是描述，而非因果检验：没有特征被消除或添加，并且缺少 Swiss-Prot 注释会被视为阴性，因此标签得分是下限。

G.4 案例研究

找出判定所依赖的残基。 图 16 展示了对四种在 UniProt 中没有活性位点注释的酶进行的单残基敲除。在预测折叠结构的区域中，分别单独敲除每个残基，以确定移除后使模型的催化判定降低最多的残基。在每种情况下，一个残基承载了大部分影响，而该残基正是酶家族的已知机制所指向的残基。读出结果也存在遗漏：在同十二个案例研究中的三种 cGAS 样核苷酸转移酶和一种腺苷酸激酶中，决定性残基是核苷酸结合环中的甘氨酸，而不是已知的催化残基。

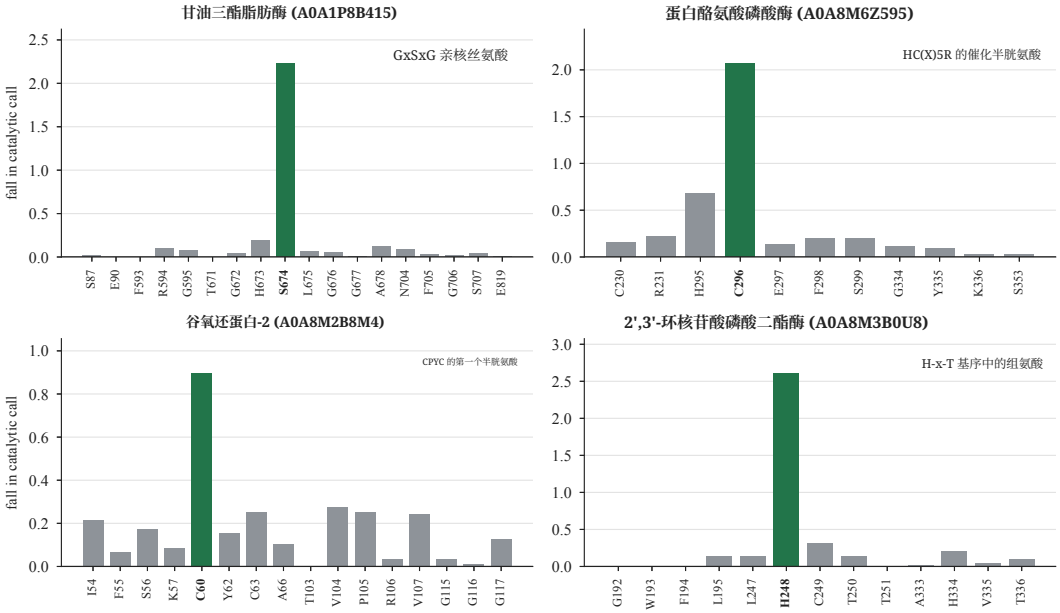


图 16：四种酶中最重要区域内的单残基敲除，这些酶均未注释活性位点。柱状图显示模型的催化判定（一个 logit）下降了多少；最大值以绿色显示。每个面板中的注释是该家族的教科书式催化残基。最终基因本体检查点。

协同作用的残基。图17比较了逐个敲除残基与同时敲除多个残基的效果。对于去泛素化酶的半胱氨酸–组氨酸二联体，联合效应超过了两者单独效应之和。两种蛋白质被注释为“可能无活性”的酶，但它们的催化残基确实存在，而且模型的判断取决于这些残基的协同作用：蛋白酶中的两个天冬氨酸残基共同敲除造成的影响，是分别敲除时影响总和的 3.4 倍。在细胞球蛋白中，模型的判断依赖于结合血红素铁的近端组氨酸，而不是远端组氨酸。这些结论描述的是模型所依赖的因素；这两种“无活性”蛋白质是否具有活性，则有待实验验证。

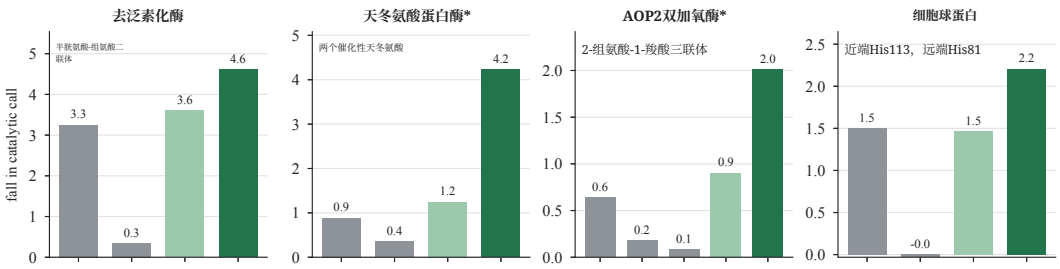


图17：单独敲除的残基（灰色）、这些效应的总和（浅绿色），以及同时敲除它们的效应（深绿色）。*在 UniProt 中标注为“可能无活性”。每种联合效应都大于30组匹配的远距离残基所产生的效应。

对 DORIP1 的一项未经验证的预测。DORIP1 与多巴胺D1受体结合，但其作为酶的作用尚未得到证实。在回答有关小鼠蛋白质的问题时，RumiFold 预测它可能是一种蛋白酶，即能切割蛋白质的酶。这项预测主要依赖 Cys57，其次依赖 His265。尽管这些残基在序列中相距甚远，但在小鼠和人类 DORIP1 的预测结构中，它们彼此都很接近（图18）。

RumiFold 的两种功能预测方式给出了不同的答案。它的文字回答指出这是一种半胱氨酸蛋白酶，而单独的功能分类器对该活性给出的评分较低。DORIP1 是否真的会切割蛋白质，仍是一个悬而未决的问题。

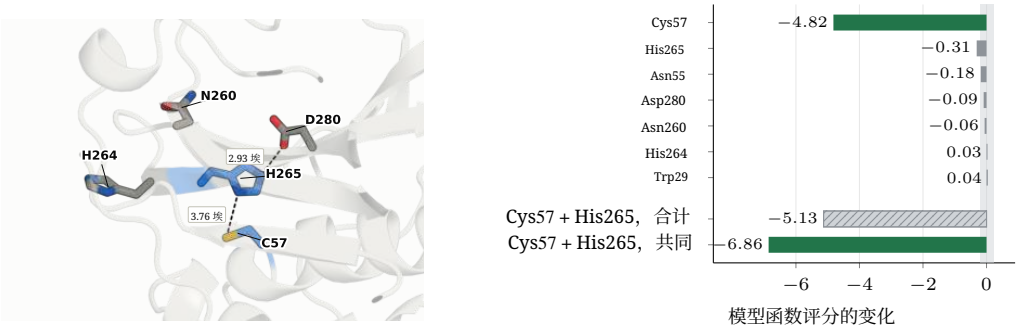


图18：**小鼠DORIP1 (Q4KL13)：一种假设，尚未经过实验验证**。*左*：预测结构（AlphaFold 数据库）中的Cys57和His265，相距3.8 埃。*右*：敲除某个残基时模型函数评分的变化。将Cys57和His265一起移除所造成的影响，大于分别单独移除二者的影响之和；测试的其他残基均落在匹配对照组的范围内（灰色带）。

H 基于证据的推理与示例答案

用模型的话说，从残基到功能。在功能训练期间，模型需要按固定顺序写出答案：它找到的功能残基及其在自身预测折叠结构中的距离；这些残基所暗示的含义；然后是从一般到具体的基因本体谱系。图19展示了留出蛋白质中的四个此类答案。前两个答案中，对于训练数据里没有已注释家族的蛋白质，残基、化学反应和功能全部正确。第三个答案中，模型找到了铁中心和正确的反应类别，但将其命名为错误的特定酶。第四个答案中，它找到了催化性天冬氨酸，随后却将该蛋白质称为丝氨酸蛋白酶：所陈述的证据与结论并不一致。在这次评估的113个蛋白质中，残基和功能均正确的有57个，均错误的有18个，另有38个只有其中一项正确。我们在第7节中表明，将这些观察结果写出来并不能改善功能预测；在G.1节中表明，模型并不会读取它所引用的数据行；这种格式可用于观察模型建立了哪些关联，但不能作为其计算方式的证据。

锌金属内肽酶（C5FUK3）：正确

模型。观察结果：活性位点：Glu432。结合 Zn(2+)：His431、His435、His441（Cα–Cα 7–10 埃，属于同一个簇）。推断：辅因子：Zn(2+)。这些残基构成一个由 2 组、共 4 个残基组成的催化位点。该化学反应为水解。功能：催化活性 ▷水解酶活性 ▷肽酶活性 ▷内肽酶活性 ▷金属内肽酶活性。EC 3.4.24.-
.
参考。活性位点：Glu432、His441。结合 Zn(2+)：His431、His435、His441、Cys623。相同的化学反应和相同的功能谱系；EC 3.4.24.-。

去泛素化酶（Q5VV17）：正确

模型。观察结果：活性位点：Asp317、Cys320、His431（Cα–Cα 5–9埃，属于同一个簇）。推断：这些残基构成一个由 3 个残基组成的催化位点。该化学反应为水解。功能：催化活性 ▷水解酶活性 ▷肽酶活性 ▷泛素样蛋白肽酶活性 ▷去泛素化酶活性 ▷半胱氨酸型去泛素化酶活性。EC 3.4.19.12。参考文献。 相同的三个残基（5–10 埃）、相同的谱系和 EC 编号。

铁依赖性双加氧酶（D0E8I4）：类别正确，酶错误

模型。观察结果：结合铁阳离子：His118、Asp120、His198（Cα–Cα 6–8 埃，形成一个簇）。推断：这些残基形成一个由 3 个残基组成的铁中心。化学反应属于氧化还原。功能：催化活性 ▷氧化还原酶活性 ▷氧化还原酶活性，作用于成对供体，并伴随分子氧的掺入或还原 ▷2-氧代戊二酸依赖性双加氧酶活性。EC 1.14.11.55。参考。 相同的残基和谱系，但 EC 1.14.11.46，即（2-氨基乙基）膦酸双加氧酶。

天冬氨酸肽酶（Q45832）：结论与证据相矛盾

模型。观察结果：活性位点：Asp174。推断：这些残基形成一个由 1 个残基组成的催化位点。化学反应属于水解。功能：催化活性 ▷水解酶活性 ▷肽酶活性 ▷丝氨酸型肽酶活性 ▷丝氨酸型内肽酶活性。EC 3.4.21.-。 参考。活性位点：Asp174。功能：.. 。 ▷ 内肽酶活性 ▷ 天冬氨酸型内肽酶活性。EC 3.4.23.-。

图19：在留出蛋白质上的回答，按照观察、推断、功能的顺序撰写，并经过轻微重新排版（已移除链标签）。这些回答来自功能训练阶段的一个检查点（第 126k 步），早于结果表中的模型。

表17：**将模型自身的残基预测放入提示中会降低功能准确率，而其内容并不重要**。GO-GPT v2 测试集（3,726 种蛋白质）上的信息加权 F1，使用与表 1 相同的评分器。除第一行外，每一行中，模型都先指出功能残基，再根据提示中的这些文本预测功能。 Δ：与直接预测相比的均值配对差异（95% 自助法区间均在 ± 0.008 以内）。证据行和 Δ 是在较早的检查点上测得的，该检查点的直接预测得分为 0.682。

提示中的证据	MF	BP	CC	Mean	Δ
无（直接预测功能）	**0.685**	**0.615**	**0.748**	**0.683**	
自身的残基预测	0.656	0.495	0.722	0.624	−0.058
自身的残基预测，仅保留置信度高的预测	0.656	0.498	0.725	0.626	−0.056
自己的自由文本观察	0.655	0.506	0.725	0.629	−0.053
对照					
自己的预测，位置随机重新分配	0.657	0.497	0.717	0.623	−0.058
另一种蛋白质的预测	0.655	0.499	0.722	0.625	

说明证据有帮助吗？ 表17检验模型能否利用自己写下的观察。结果是不能。各种形式的证据都会使加权平均F1降低 0.053 至 0.058，其中几乎全部损失都出现在生物过程上（-0.117；分子功能为 -0.029，细胞组分为 -0.028）。模型自身的残基预测并不比位置随机重新分配后的相同陈述（$+0.001$，$[-0.003, +0.005]$）或另一种蛋白质的预测（-0.001，$[-0.006, +0.004]$）更好：模型对文本是否出现有反应，却不关注文本表达的内容。与训练集中存在近缘蛋白质的蛋白质相比，这种损失最大（序列一致性达到 90% 或更高时为 -0.083，低于 30% 时为 -0.030），而我们考察的任何子集都未显示出可靠的提升。

对其从未见过的蛋白质所作的自由文本描述。图20展示了为新颖的 Prot2Text 测试蛋白质撰写的描述，这些蛋白质的参考文本未出现在训练集中。典型的

错误见于中间三个案例：模型正确识别了蛋白质的类型及其作用，但在一个具体细节上出错，例如底物或互作伙伴。最后一个则是彻底失败。

P0A625：正确

模型。 钼结合蛋白依赖型转运系统的一部分；可能负责底物跨膜转位。**参考。** 钼结合蛋白依赖型转运系统ModABCD的一部分；可能负责底物跨膜转位。

Q7S133：作用正确，互作伙伴名称错误

模型。 SWR1复合物的催化组分，可介导组蛋白H2A与H2A变体HZT1之间的ATP依赖性交换，并通过染色质重塑调控特定基因的转录。**参考。** .....组蛋白H2A与H2A.Z之间的交换，并通过染色质重塑调控特定基因的转录。

P0AE25：机制正确，底物错误

模型。 半乳糖通过边界膜摄入，同时质子进入细胞（同向转运系统）。 **参考。** L-阿拉伯糖通过细胞质膜摄入，同时质子进入细胞（同向转运系统）。

O31266：正确的酶类别，错误的底物

模型。 参与氧代喹啉降解和利用的开环双加氧酶。 **参考。** 参与喹哪啶降解和利用的开环双加氧酶。

O13563：错误

模型。 与副肌球蛋白α链的共识序列相匹配。**参考。** 19S帽蛋白酶体复合物的组成成分，作为26S蛋白酶体的调节亚基，参与ATP依赖性的泛素化蛋白质降解。

图20：为五个Prot2Text新测试蛋白质生成的描述及参考描述（RumiFold在第240k步，采用贪心解码）。

被问及自身时，模型回答的是蛋白质相关内容。 为了非正式地探究语境中的蛋白质如何为日常对话增添色彩，我们向基因本体模型提供了一种蛋白质，并提出了一个与其毫无关系的问题：“你接受了什么训练？”图21展示了模型的回复。答案会随蛋白质而变化。面对作为肽酶的ACE2时，模型仿佛自己就是这种酶，并列出了其底物；面对胰岛素时，它说自己接受过护士培训；面对Tau蛋白和p53时，它描述了医学教育经历。在20多种蛋白质中，有8条回复描述了医学或临床培训，6条是语言模型的套话式回答，还有一条针对T4溶菌酶的回复列出了蛋白质序列和结构数据，这与事实最为接近。这些都是单次贪心采样，标签由人工分配，因此只是轶闻，并非经过测量的效应。我们之所以纳入这些例子，是因为它们显示出折叠记忆延伸到了模型从未接受训练去回答的内容中；这正是第7节所报告的对通用语言能力造成的代价的一种表现。

ACE2 (Q9BYF1)，一种肽酶：以酶的身份作答

模型。 我接受了多种底物的训练，其中包括但不限于蛋白质、肽以及其他膳食材料。我的训练涉及学习如何识别并响应不同类型的刺激，例如识别待加工的目标（如肽或蛋白质），然后根据该目标的要求对其进行加工。

胰岛素（P01308）：临床角色

模型。 我接受过护士训练。

Tau蛋白（P10636）：医学课程体系

模型。 我接受过全面的课程体系训练，涵盖生物学、解剖学和生理学等基础知识，之后还包括医学、临床技能和职业伦理等方面的专业教学。我的训练旨在帮助我胜任医学职业，重点培养批判性思维、精准度和以患者为中心的护理。

p53（P04637）：肿瘤学优先

模型。 我的训练数据非常庞大，包括但不限于医学文献、研究论文和真实世界的临床情境。我的训练基于来自不同领域的信息，例如肿瘤学、心脏病学、药理学等。。。

T4溶菌酶（P00720）：最接近真相

模型。 我的训练数据非常庞大，包括但不限于：基因组序列（DNA、RNA 等）；
蛋白质及其结构域的结构信息；分类学信息；功能（例如酶底物、
反应机制）；指示特定结合或加工过程的序列特异性模式。。。。

图21：在不同蛋白质的上下文中，对“你接受过哪些训练？”的回答，内容来自 RumiFold 的基因本体微调（贪心解码；每种蛋白质取一个样本；T4溶菌酶的回答由列表精简而成）。